Journal of Official Statistics
vol. 39, 3 (September 2023)

Journal of Official Statistics, Vol. 39, No. 3, 2023, pp. 275–284, http://dx.doi.org/10.2478/JOS-2023-0013

Letter to Editor
Quality of 2017 Population Census of Pakistan by Age and Sex

Asif Wazir[1] *and Anne Goujon*[2]

This Letter to Editor is a supplement to the previously published article in the Journal of Official Statistics (Wazir and Goujon 2021).

In 2021, a reconstruction method using demographic analysis for assessing the quality and validity of the 2017 census data has been applied, by critically investigating the demographic changes in the intercensal period at national and provincial levels. However, at the time when the article was written, the age and sex structure of the population from the 2017 census had not yet been published, making it hard to fully appreciate the reconstruction of the national and subnational level populations.

In the meantime, detailed data have become available and offer the possibility to assess the reconstruction's outcome more in detail. Therefore, this letter aims two-fold: (1) to analyze the quality of the age and sex distribution in the 2017 Population census of Pakistan, and (2) to compare the reconstruction by age and sex to the results of the 2017 population census. Our results reveal that the age and sex structure of the population as estimated by the 2017 census suffer from some irregularities. Our analysis by age and sex reinforces the main conclusion of previous article that the next census in Pakistan should increase in quality with an inbuild post-enumeration survey along with post-census demographic analysis.

Key words: Census quality; population reconstruction; demographic analysis; Pakistan, Pakistani provinces.

1. Quality of 2017 Population Census of Pakistan by Age and Sex

The constitution of Pakistan does not require the census to be held every ten years as, for example, the U.S. Constitution does. The ten-year period is an international convention under which Pakistan held the census till 1981. Thereafter Pakistan held only two censuses in the last 40 years, that is, in 1998 and 2017. Like in many settings, the censuses in Pakistan have been subject to various types of non-sampling errors including misreporting, nonresponse, recall bias, etc. These quality concerns have not been properly addressed as the Pakistan Bureau of Statistics did not implement a post-enumeration survey (PES) for the last four censuses (though a PES was conducted for the 1981 census, however, results were not published) which was an essential component to assess the potential coverage and content errors in the census enumerations. The census that was held in 2017 was no exception.

[1] United Nations Population Fund, Maputo, Mozambique. Email: wazir@unfpa.org

[2] Population and Just Societies Program, International Institute for Applied Systems Analysis, Schlossplatz 1, 2361 Laxenburg, Austria. Email: goujon@iiasa.ac.at

Acknowledgments: This study is based on publicly available data and was not funded. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the United Nations Population Fund (UNFPA). The authors declare no conflict of interest.

In 2021, we applied, for the first time in the case of Pakistan, a population reconstruction method for assessing the quality and validity of the 2017 census data, by critically investigating the demographic changes in the intercensal period at national and provincial levels, using a range of intercensal surveys (Wazir and Goujon 2021). We showed that on the one hand, while the reconstructed population estimates at the national level can be seen as broadly valid, as compared to the 2017 census figures, on the other hand, the reconstruction does not perform so well at the sub-national level, notably because of the uncertainty involved around internal migration between 1998 and 2017.

However, at the time when the article was written and was under consideration by the Journal of Official Statistics, the age and sex structure of the population from the 2017 census had not yet been published, making it hard to fully appraise the reconstruction of the national and regional populations. In the meantime, detailed data have become available and offer the possibility to assess the reconstruction's outcome in more detail.

The aim of this letter is two-fold: (1) to analyze the quality of the age and sex distribution in the 2017 Population census of Pakistan, and (2) to compare the reconstruction by age and sex to the results of the 2017 population census.

2. Evaluation of Data Quality of the 2017 Population and Housing Census of Pakistan

As a starting point, the single-year age and sex distribution are examined visually to check for anomalies. Figure 1 shows the single-year age distribution by sex for the 1998 and 2017 censuses.

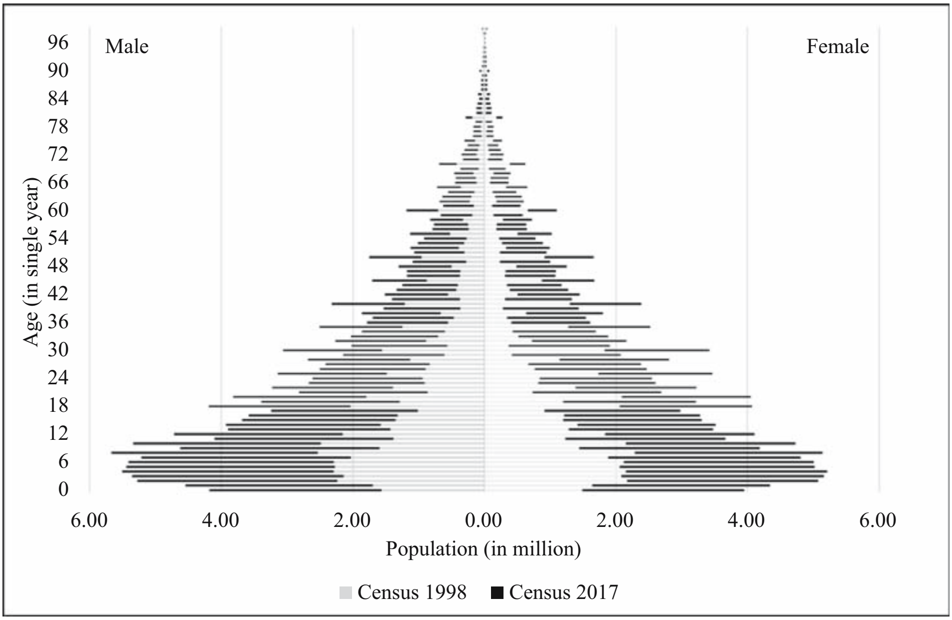


Fig. 1. Single-year age and sex distribution from the 1998 and 2017 censuses in Pakistan.
Source: Authors' calculations based on single-year age-sex data (Pakistan Bureau of Statistics 2022; Minnesota Population Center 2023).

The combined pyramids for the 1998 and 2017 censuses indicate that irregularities in the age distribution are significant for both men and women. The last two censuses show extensive age heaping for males and females alike, with the usual concentration around numbers ending in 0 and 5. Significant under-reporting of children (age $<$ ten years) and age misreporting among the working-age population are visible for both sexes in both censuses. The distortions are less common among older ages. It is worth noting, that their magnitude is less in the 2017 census compared to the 1998 census, suggesting improvements in the accuracy of the age data. This is mainly because the 2017 census used the national identity card (as well as the verified computerized national identity card from the central database) for the census questionnaire enumeration. Therefore, the date of birth was reported more precisely in the 2017 census.

3. Sex Ratio by Age

In the 2017 census, the share of men in the total population was 51.2% at the national level. The proportion was higher in Balochistan (52.6%) and lower in Punjab (50.8%). While the proportion of men in Sindh and KP stood at 52% and 50.6%, respectively.

Further, we also illustrate the sex ratio by single-year age group from the 1998 and 2017 censuses in Figure 2. The age heaping and misreporting in the 1998 census (shown in Figure 1) manifested itself in the distorted pattern of the sex ratio by age, at all ages, decreased in the 2017 census. The sex ratio at birth as well as up to the age of 15 showed a preference for boys in both censuses in Pakistan. Advances in sex-selective abortion technology and persistent discrimination in care practices for girls have resulted in higher female mortality. Between ages 20 and 40, the sex ratio reversed mainly due to slightly higher male mortality at younger ages, and potentially male international emigration for employment, particularly in the gulf countries. Estimates from the Bureau of Emigration and Overseas Employment (2020) show that 5.1 million people migrated overseas for

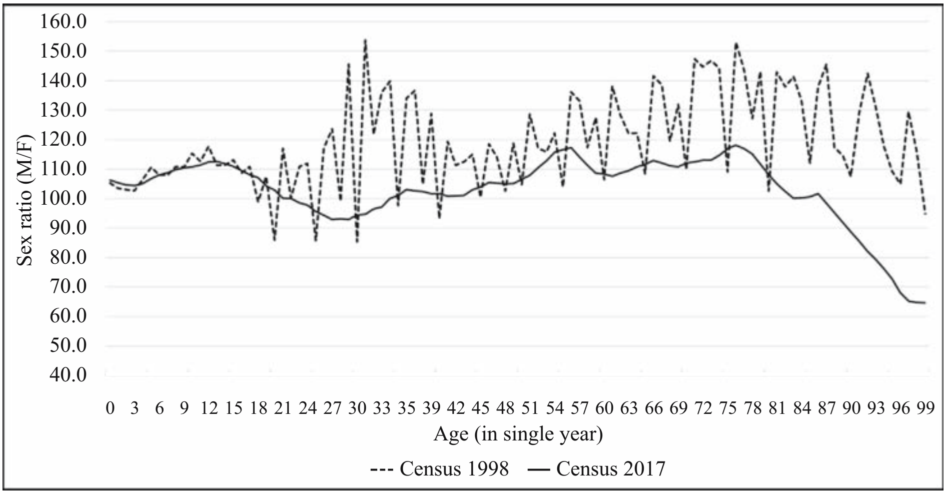


Fig. 2. Sex ratio by single year of age from the 1998 and 2017 censuses in Pakistan.
Source: Authors' calculations based on single-year age-sex data (Pakistan Bureau of Statistics 2022; Minnesota Population Center 2023).

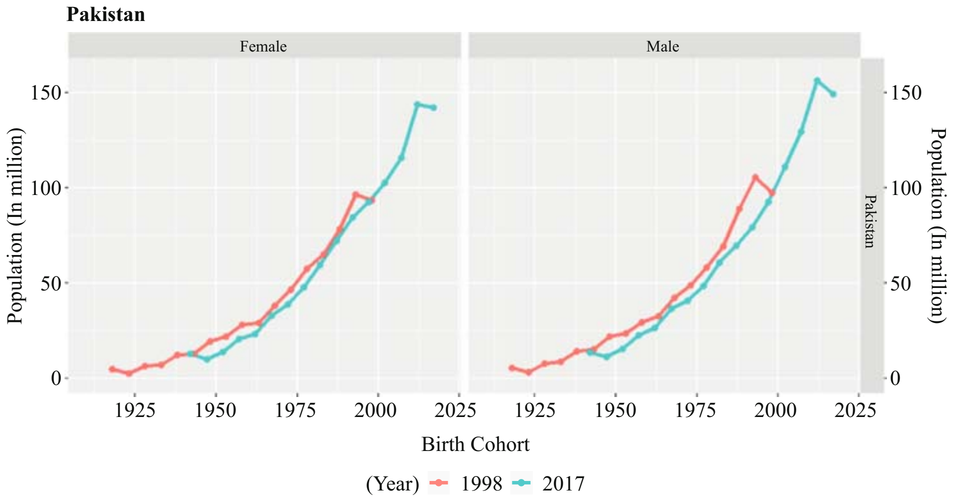


Fig. 3. Birth cohort age-sex distribution from the 1998 and 2017 censuses in Pakistan.
Source: Authors' calculations based on single-year age-sex data (Pakistan Bureau of Statistics 2022; Minnesota Population Center 2023).

employment through formal channels between 2010 and 2017. Finally, at older ages, the sex ratio continuously dropped below 100 due to the female advantage in mortality.

We also performed a graphical birth cohort analysis based on the population data from the last two censuses for men and women in Pakistan depicted in Figure 3. In ideal circumstances, the size of each birth cohort should decline over time (1998 and 2017) due to the mortality, with the assumption that the country did not experience significant migration. In the case of Pakistan, three important patterns emerged from Figure 3: first, the significant under-enumeration of boys and girls under the age of five in the censuses are evident; second, the significant differences in the patterns between the male and female birth cohorts at young ages reinforcing the under-enumeration for both male and female; and third, the almost overlapping of birth cohorts between the two censuses, especially for women. These inconsistencies by age and sex confirmed the presence of irregularities in the enumeration process, particularly in the 2017 census.

Apart from the visual check, we have computed the Whipple's Index (Siegel and Swanson 2004) and the Myers Index (Myers 1954), using the Population Analysis System (PAS) software developed by the U.S. Census Bureau (U.S. Census Bureau 2017). While at the national level and for larger regions like Punjab and Sindh, the 2017 census indices show fair results (in terms of age heaping and/or digit preference for both sexes), small regions have issues of inaccuracy in age reporting, particularly Balochistan and Khyber Pakhtunkhwa (KP).

4. Age Ratio Score

The age ratio score which measures the divergence of one age group from the neighboring two age groups is a valid indicator of errors in age reporting. In the absence of significant changes in fertility, mortality, and migration, the enumerated size of a particular cohort should be approximately equal to the average size of the preceding and following cohorts.

Without irregularities, the age ratio score should be approximately equal to 100. For example, Gerland (2014) utilized the age ratio scores to assess the quality of the census data in India. Figure 4 indicates the age-ratio scores for males (Figure 4a) and females (Figure 4b) from the 1998 and 2017 censuses. The age-ratio scores also reinforced the obvious irregularities in all age groups in both censuses: over-enumeration in the 5–9 age group, and under-enumeration in the 10–14 age group for both boys and girls. For age groups ranging from 15–19 to 45–49, the irregularities were noticeable for men and women in the same way, with higher intensity in the 1998 census. From age 50 onwords, substantial anomalies became evident for both sexes in both censuses, which may be related to age misstatement/exaggeration and the mortality patterns in older persons. Although age-reporting errors have persisted between the last two censuses, the 2017 census shows marked improvements compared with the 1998 census.

Based on the age ratio scores, we have calculated the age-sex accuracy index for five-year age groups (Siegel and Swanson 2004), which is the sum of the age ratio score for

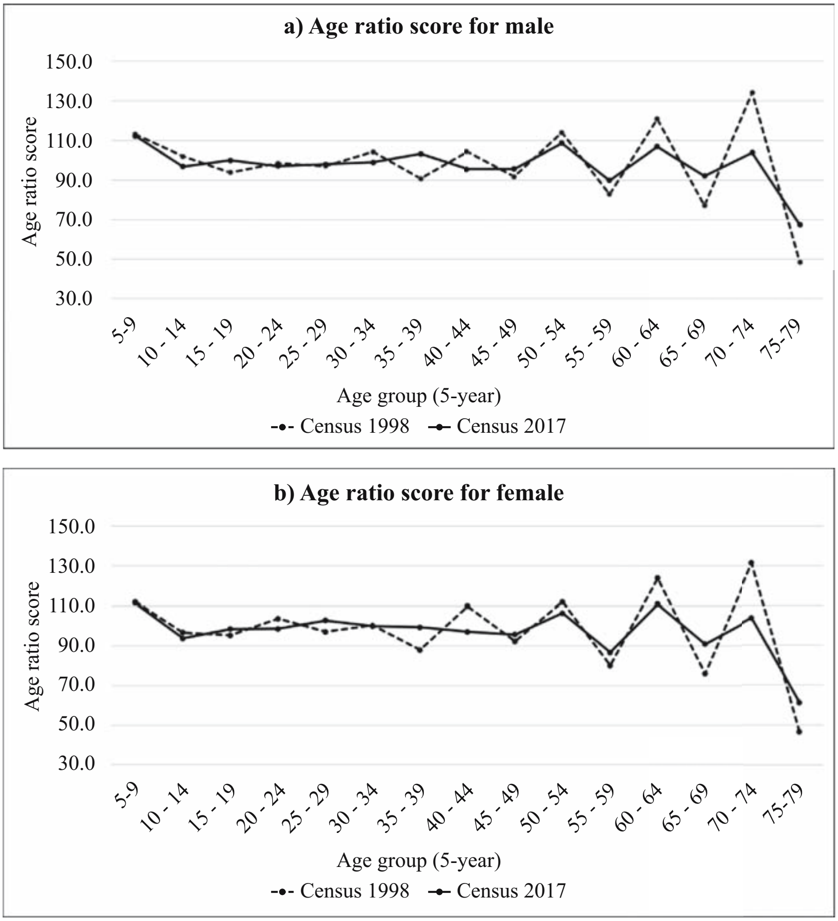


Fig. 4. Age ratio score by age and sex in Pakistan, 1998 and 2017 population censuses.
Source: Authors' calculations based on age-sex data (Pakistan Bureau of Statistics 2022: Minnesota Population Center 2023).

Table 1. Age-sex accuracy index, four provinces, Islamabad, and Pakistan, 2017.

	Index
Balochistan	31.9
KP	25.9
Punjab	22.2
Sindh	25.4
Islamabad	29.5
Pakistan	22.0

Source: Authors' calculations based on 2017 population census data (Pakistan Bureau of Statistics 2022).

males, the age ratio score for females, and three times the sex ratio score. The accuracy scores are divided into three ranges to assess the magnitude of the reporting problem: (1) when the accuracy index is $<$ 20, the census estimates are accurate; (2) between 20 and 40 the estimates are inaccurate, and (3) the estimates are highly inaccurate when the score is $>$ 40. To compute the age-sex accuracy index, we used the "ageSexAccuracy" function from the R package "DemoTools" (Riffe et al. 2019) and applied it to the national level as well as to the four provinces and Islamabad (Table 1).

The age distribution of the 2017 population census at the national level falls under the inaccurate ($>$20) categories. However, there is a large variation observed at the sub-national level. All provinces were also categorized as inaccurate, the lowest value of the index was found in Punjab (22.2) and the highest in Balochistan (31.9). The analysis shows that the distribution by age and sex resulting from the 2017 census was somehow distorted at the sub-national level.

5. Comparison of the Age-Sex Distribution According to the 2017 Census and the Reconstruction

In 2021, we used demographic analysis to estimate the population size by age and sex based on inter-census (1998–2017) estimates of births, deaths, and internal and international migration (Wazir and Goujon 2021). Here, we compare the reconstructed 2017 population with the official results of the 2017 census population by age and sex using a goodness-of-fit-statistical test computed as the absolute difference (official minus reconstructed population) and the average absolute deviation by age and sex (see Figure 5) along with the 95% confidence interval (CI). An overcount refers to the census population being higher than the reconstructed population, while an undercount refers to the opposite situation.

While our analysis of the total population showed little difference at the national level, we find that there are quite substantial differences by age and sex. We find an overcount of children under age 15 in the census and an undercount for all other age groups (from 15-19 to 75+). There were ten million (95% CI from 3.0 to 15.5) more children under the age of 15 in the census compared to the reconstruction at the national level, with 5.5 million (95% CI from 4.4 to 10.4) boys and 4.5 million (95% CI from 3.9 to 9.3) girls.

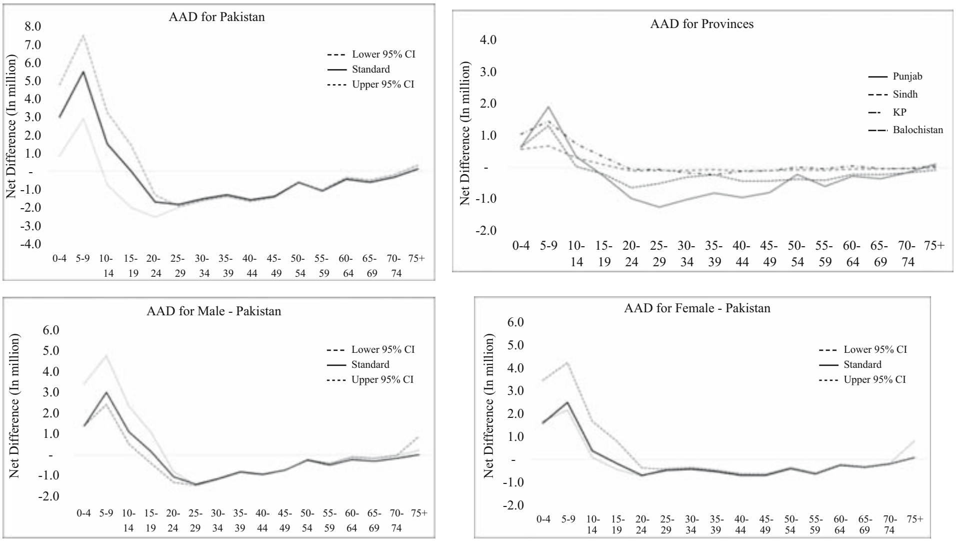


Fig. 5. Average absolute deviation (AAD) by age and sex at the national and provincial level.
Source: Authors' calculations based on 2017 population census data (Pakistan Bureau of Statistics 2022; Wazir and Goujon 2021).

At the sub-national level, the overcount in the 0–14 age group was spread as such: 3 million in Punjab (95% CI from −0.1 to 5.1) and 2.1 million in Sindh (95% CI from −0.4 to 3.4). The overcount was substantial in KP with 3.3 million (95% CI from 1.8 to 4.7) and in Balochistan with 1.6 million (95% CI from 0.9 to 2.2), which are relatively small regions. It is worth noting that the difference was particularly acute for the 5-9 age group compared to the 0–4 and 10–14 ones (see Figure 5 and Table 2).

Table 2. Average absolute deviation by age at the national and provincial level (in millions).

	Punjab	Sindh	KP	Balochistan	Pakistan
0–4	+0.7	+0.7	+1.1	+0.6	+3.0
5–9	+1.9	+1.3	+1.5	+0.7	+5.5
10–14	+0.4	+0.1	+0.8	+0.3	+1.5
15–19	−0.3	−0.2	+0.3	+0.1	+0.0
20–24	−1.0	−0.6	+0.0	−0.1	−1.7
25–29	−1.2	−0.5	+0.0	−0.1	−1.8
30–34	−1.0	−0.3	−0.1	−0.1	−1.5
35–39	−0.8	−0.2	−0.2	+0.0	−1.3
40–44	−0.9	−0.4	−0.1	+0.1	−1.6
45–49	−0.8	−0.4	−0.1	+0.1	−1.4
50–54	−0.2	−0.4	+0.0	−0.1	−0.6
55	−0.6	−0.4	+0.0	−0.1	−1.1
60–64	−0.3	−0.2	+0.1	+0.0	−0.4
65–69	−0.3	−0.2	+0.0	+0.0	−0.6
70–74	−0.1	−0.1	+0.0	+0.0	−0.3
75+	+0.1	−0.1	+0.1	+0.0	+0.1

Source: Authors' calculations based on 2017 population census data (Pakistan Bureau of Statistics 2022; Wazir and Goujon 2021).

The difference in the age structure between the census and the reconstruction can have several causes. Besides the deficiencies of the census as noted in the previous section, it is possible that the fertility and mortality levels used in the reconstruction and originating from the existing demographic surveys underestimate the number of births and deaths, especially at the sub-national level.

6. Comparison of the Sex Ratio According to the 2017 Census and the Reconstruction

Figure 6 shows the sex ratio (number of men to 100 women) for all ages according to three sources: the 2017 population census, model life tables, and the reconstruction. The sex ratio computed from the model life table East Model life table (Coale et al. 1983), is based on the life expectancies at birth for provinces and national level (as calculated by Wazir and Goujon 2021). The model life tables show that sex ratios should be declining with age, which was not the case in the reconstruction. Except for KP, the sex ratios for children up to age 15 from the reconstruction were close to the value computed from model life tables, while the ratios calculated from the 2017 census show an irregular pattern, for all ages. The sex ratios from the reconstruction between ages 15 and 50 were smooth and matched those from the model life tables, except for Sindh and Balochistan. In these two provinces, females were likely to be underreported, leading to systematic excess of males in the younger cohorts.

The sex ratios for older ages (age 50 and above) estimated from the census and the reconstruction were contradictory to those using model life tables for all provinces and the national level. All provinces and national levels experienced a significantly high level of sex ratios – above 100 – for older ages and remained persistent. The plausibility of these distortions in older ages is difficult to explain, whether they were due to higher mortality of women, systematic under-reporting of older cohorts, or other omission issues. Those

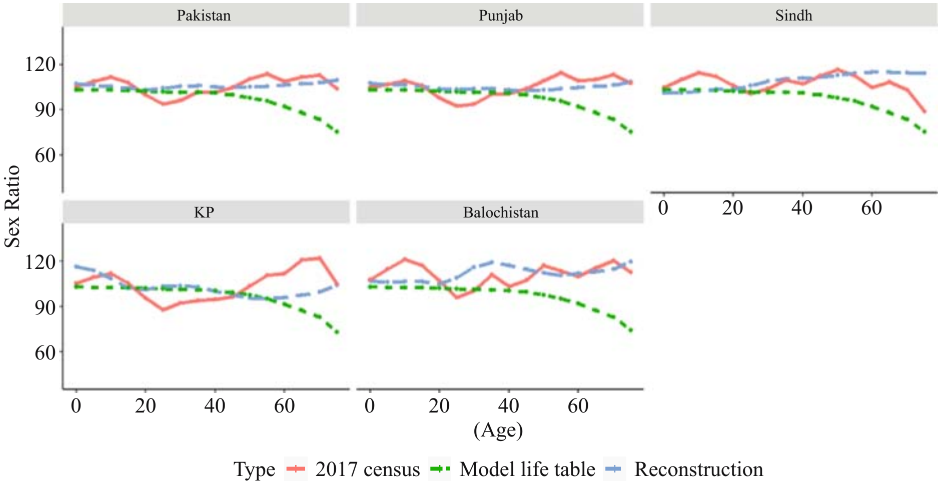


Fig. 6. Sex ratio by age in Pakistan from 2017 census, model life table, and reconstructed population.
Source: Authors' computation based on 2017 population census data (Pakistan Bureau of Statistics 2022; Coale et al. 1983; Wazir and Goujon 2021).

results tend to point in the direction of data issues in the 2017 census as well as in the mortality and internal migration indicators estimated and used in the reconstruction.

Finally, the inconsistencies in the pattern of sex ratios in adulthood in the 2017 census suggest that this was a real structural anomaly, potentially representing reporting inaccuracies. Therefore, an in-depth province-specific investigation would be necessary to disentangle these factors.

7. Conclusion

Our results reveal that the age and sex structure of the population as estimated by the 2017 census suffered from some irregularities, as is the case in many low-income countries, both at the national and sub-national levels. These irregularities partly explained the observed differences between the census and the reconstruction. Similarly, the validity of the reconstruction is limited by the data estimates, in terms of fertility, mortality, and internal migration, which become more visible when looking at the age structure. At the sub-national level, the overall irregularities were amplified by issues related mostly to the difficulties in estimating internal migration. This was particularly visible in the small provinces of Balochistan and KP.

The analysis by age and sex reinforces the main conclusion of our article that the government and the international community should ensure that the next census in Pakistan increase in quality with an inbuild post-enumeration survey along with post-census demographic analysis. To resolve the contention about the provincial population size, the census should use both dejure and de facto population count within a short enumeration period (ideally three to five days) and collect the information on the previous place of residence and duration of stay at the current residence (status of internal migration).

8. References

Bureau of Emigration and Overseas Employment. 2020. Labor Migration Report. Islamabad: Bureau of Emigration and Overseas Employment. Available at: https://beoe.gov.pk/reports-and-statistics.

Coale, A., P. Demeny, and B. Vaughan. 1983. Regional Model Life Tables and Stable Populations (Vol. 2nd Edition). New York 10003, NY, USA: Academic Press Inc.

Gerland, P. 2014. UN *Population Division's Methodology in Preparing Base Population for Projections: A case study for India* Asian Population Studies: 10(3): 274–303. DOI: https://doi.org/10.1080/17441730.2014.947059.

Minnesota Population Center 2023. *Integrated Public Use of Microdata* Series (IPUMS), International: Version 6.5 [data set]. Minneapolis, MN: University of Minnesota.

Myers, R.J. 1954. "Accuracy of Age Reporting in the 1950 United States Census." *Journal of the American Statistical Association* 49(268): 826–831. DOI: https://doi.org/10.2307/2281542.

Pakistan Bureau of Statistics. 2022. *Final Results of 6th Population and Housing Census*, Islamabad: Pakistan Bureau of Statistics, Ministry of Statistics, Islamabad, Islamic

Republic of Pakistan. Available at: https://www.pbs.gov.pk/content/final-results-census-2017-0 (accessed January 2023).

Riffe, T., J.M. Aburto, M. Alexander, S. Fennell, l. Kashnitsky, M. Pascariu, and P. Gerland. 2019. *DemoTools: An R Package of Tools for Aggregate Demographic Analysis* (version 01.13.76). Available at: https://github.com/timriffe/DemoTools/ (accessed January 2023).

Siegel, J., and D. Swanson. 2004. *The Methods and Material of Demography*. San Diego, California, USA: Elsevier Academic Press. Available at: https://books.emeraldinsight.com/page/detail/the-methods-and-materials-of-demography/?k=9780126419559.

U.S. Census Bureau. 2017. *Population Analysis System (PAS) Software*. Available at: https://www.census.gov/data/software/pas.html (accessed January 2023).

Wazir, A., and A. Goujon. 2021. “Exploratory Assessment of the Census of Pakistan Using Demographic Analysis.” *Journal of Official Statistics* 37(3): 719–750. DOI: https://doi.org/10.2478/jos-2021-0032.

Received July 2022
Revised November 2022
Accepted February 2023

Journal of Official Statistics, Vol. 39, No. 3, 2023, pp. 285–324, http://dx.doi.org/10.2478/JOS-2023-0014

Looking for a New Approach to Measuring the Spatial Concentration of the Human Population

Federico Benassi[1], *Massimo Mucciardi*[2], *and Giovanni Pirrotta*[3]

In the article a new approach for measuring the spatial concentration of human population is presented and tested. The new procedure is based on the concept of concentration introduced by Gini and, at the same time, on its spatial extension (i.e., taking into account the concept of spatial autocorrelation, polarization). The proposed indicator, the Spatial Gini Index, is then computed by using two different kind of territorial partitioning methods: MaxMin (MM) and the Constant Step (CS) distance. In this framework an ad hoc extension of the Rey and Smith decomposition method is then introduced. We apply this new approach to the Italian and foreign population resident in almost 7,900 statistical units (Italian municipalities) in 2002, 2010 and 2018. All elaborations are based on a new ad hoc library developed and implemented in Python.

Key words: Spatial concentration; Gini index; constant step; maxmin distance; Italy.

1. Introduction

In population studies and, more generally, in the quantitative social sciences, concentration, and thus space, assumes a fundamental importance (Anselin 1999; Logan 2012; Howell et al. 2016). Indeed, people and firms tend to concentrate in space almost naturally, both to facilitate interactions, exchange ideas, goods and services, and to share the costs associated with survival itself. It is no coincidence that Aristotle, when defining man as a social animal ("zoon poolitkon"), refers to the concept of the arena, that is, space, and that he places this character of the individual at the birth of the polis (i.e., the city, concentrated in space, by definition). Population and space are therefore two closely interconnected and mutually dependent variables. Indeed, as Livi Bacci (1999) reminds us, not unlike other living species, humans need space to obtain the resources necessary for their survival, to maintain population growth and to organise themselves socially. On the other hand, space itself "depends" on human behaviour, not only in relation to its negative externalities, but also to the capacity to absorb human activity, the so-called "carrying capacity" (Verhulst 1838; Pearl and Reed 1920). From a more strictly statistical point of view, concentration is something intimately connected to the concept of variability, that is, the ability of a quantity to assume different values.

The statistical approach to measuring concentration is in fact essentially based on the concept of variability. But, as observed by Leti (1983), in reality the original concept of

[1] University of Naples Federico II, Deparment of Political Sciences, Via L. Rodinò 22, 80138, Naples, Italy. Email: federico.benassi@unina.it
[2] University of Messina, Department of Cognitive Science, Education and Cultural Studies, Messina, Italy. Email: massimo.mucciardi@unime.it
[3] University of Messina, IT Staff, Messina, Italy. Email: giovanni.pirrotta@unime.it

variability is, under certain conditions, entirely analogous to that of heterogeneity and homogeneity, as well as, to that of concentration.

In fact, for quantity distributions, any index of variability is also an index of concentration, and indeed indices of homogeneity, applied to quantitative characters, actually measure concentration (Novelli and Ocelli 1999). In this perspective, therefore, concentration would be nothing more than the variability associated with a transferable variable. On the basis of this approach to the study of concentration, which could be considered to be statistical, and which is aspatial, a vast literature was born and has developed. In this framework, one of the cornerstones is Gini's *G* index (Gini 1912, 1914, 1921, 2005), which in turn was based on the pre-existing Lorenz concentration curve (Lorenz 1905). This index has stimulated a growing school, especially among Italian statisticians and demographers, who have proposed many other indices of this type, that is, aspatial, for the study of concentration. One recalls among these, Bonferroni's index (Bonferroni 1938), and those of Zenga (1984, 1985) to which many others can be added as can be seen in Frosini (1996). But also in the international sphere, the work of Gini stimulated the definition of a number of indices, such as Wright's index (Wright 1937) and the well-known Hoover index (Hoover 1941).

However, the study of concentration according to an exclusively statistical approach has attracted increasing criticism over time. In particular, the major criticism levelled at these indices is that being aspatial indices, they deny the essentially spatial nature of the concentration process (Arbia 2001; Dawkins 2004, 2006). In essence, they do not consider the dimension of polarisation, that is, the geographical component of concentration that finds its analytical formulation in Tobler (1970) and the so called first law of geography. This lack assumes particular relevance in the measurement of all those processes, such as for instance residential segregation, where space is a foundational component (Reardon and O'Sullivan 2004). To make up for this shortcoming, spatial extensions of the concentration indices and the Gini index in particular have been proposed. Mention should be made here, among others, of those by Arbia and Piras (2009), Rey and Smith (2013), Crespo and Hernandez (2020), and Panzera and Postiglione (2020). Even more recently, Mucciardi and Benassi (2023), developing seminal ideas of Alleva (1987), Mucciardi and Bertuccelli (2007) and Mucciardi (2008b), have proposed an approach for measuring concentration by means of a spatial extension of the Lorenz curve that allows the definition of a spatial version of the Gini index.

In the present article a new approach for measuring spatial concentration of a human population is proposed and tested. The approach is partially based on that of Mucciardi (2008b) and Mucciardi and Bertuccelli (2007). Here, stress is placed on the effects of different neighbourhood structures on the computation of the Spatial Gini Index (SGI). In particular, two types of distances are proposed to determine the spatial structure of the territory: MaxMin (MM) and Constant Step (CS). Finally, an extension of the decomposition method of Rey and Smith (2013) is here proposed in the framework of SGI with MM and CS. The empirical approach is referred to Italian and foreign population resident in the Italian municipalities in 2002, 2010 and 2018. It is important to stress that given the huge number of statistical units (about 7,900) we have developed a specific computational procedure, writing an ad hoc library in Python.

This article is structured as follows. In the next section the Spatial Gini Index (SGI) is presented and discussed. Attention is paid to the way the territory can be partitioned. In this context two distance methods are presented: MM and CS. SGI is therefore presented as a measure to be built in this framework. Section 3 is devoted to the implementation of the decomposition method of Rey and Smith (2013) in the general framework of SGI and in the two decomposition methods here proposed. Section 4 presents the empirical results. The last section presents a discussion, conclusion and future developments. In the two appendices a description of the technical details of SGI with the two distance methods and the ad hoc library written in Python are provided.

2. The Spatial (Intrinsic) Dimension of Concentration and the Spatial Gini Index

Common to all spatial approaches is the recognition that, in the analysis of any phenomenon, consideration of the spatial dimension requires an adequate description of the spatial variability of the phenomenon itself (Matthews and Parker 2013). Stating that a phenomenon or a relation manifests spatial variability is equivalent to saying that that phenomenon or that relation is not spatially stationary. The presence of random fluctuations, the existence of differentiations in the perceptions and behaviour of individuals, the incompleteness or imprecise specification of the descriptive model assumed, are some of the possible causes of spatial non-stationarity (Fotheringham 1997). In consideration of this aspect, this new proposal of the Spatial Gini Index (SGI) is based on comparing how the contribution in terms of “connectivity” and “variability” varies as the geographical distance between spatial units increases. So, if the variable observed is not dependent on space, the variations between the connectivity and variability components should not differ much from each other. Before showing the construction of the SGI, in the next section we present the system of spatial weights used to determine the contribution of connectivity in spatial terms.

We believe that there are two major motivations for the implementation (first) and (after) the use of SGI by official statistics. The first major motivation is a general one, but, in our view, extremely relevant. We all know that space is a fundamental dimension to better grasp socio economic and demographic phenomena and processes. Referring only to population issues (for sake of simplicity) we know that modern demography is essentially a spatial social science (Voss 2007). The importance of space in the field of official statistics is also underpinning by Eurostat (see, for example Eurostat 2015). So, from that, second major motivation, it is quite surprisingly that, to the best of our knowledge, National Statistical Institutes (NSIs) don’t have an “official” or at least a common measure of spatial concentration. In the measure of the concentration of the Gross Domestic Product for example, NSIs and other International Institutions like the United Nations or World Bank use the G index, but this is not true in terms of measuring the spatial dimension of concentration (that we proved in the article to be a fundamental dimension of concentration). So, in our view, there is a double trouble: (1) space is relevant for measuring process and phenomenon and especially concentration but (2) there aren’t any “official” or at least commonly used measure (like in the case of G). This is why we proposed SGI. The methodology here proposed is simple and completely transparent (we didn’t build a black box) and can therefore easily implemented and replicated by NSIs and other Institutions.

2.1. Partitioning the Territory: The “MaxMin” And “Constant Step” Distance Methods

According to the Lorenz curve approach (Lorenz 1905), we need a system that can quantify the contribution of connectivity in spatial terms. Our proposal is to consider buffer or threshold distances capable of progressively creating partitions of the territory (or territorial subsets). These partitions identify neighbouring and non-neighbouring units such that each partition is disjoint from the others and the sum of all the elements of all the partitions coincides with the number of all the possible pairs between the n spatial units.

2.1.1. Partitioning With MaxMin Distance Method

As we know, the Gini index is geometrically based on the Lorenz curve (Lorenz 1905). For a sample of n dimension, the curve takes the cumulative percentile of the n units (for example individuals) on the x-axis and cumulative percentile of the variable (for example income) on the y-axis. The idea of the Lorenz curve is very simple. Given a sample of n ordered units, it’s a graph that compares the distribution of a variable with a hypothetical uniform distribution of that variable (the original contribution can be found in Lorenz 1905). Perfecting this graph would be a diagonal line at a 45° angle from the origin (meeting point of the x and y axis), indicating the population’s perfect variable distribution (line of absolute equality).

To satisfy these conditions we use the MM method (Mucciardi 2008a) that we recall below.

Suppose we have n spatial units $u_1 \ldots\ldots u_n$ in which we observe $x_1 \ldots\ldots x_n$ data and let $\boldsymbol{E^0}$ be a $n \times n$ matrix of the Euclidean distances between these units such that $d_{ij}^0 = \|u_i - u_j\|_2$ (with $d_{ij}^0 \in E^0$, $i = 1 \ldots n j = 1 \ldots n$), where $\|\cdot\|_2$ is the Euclidean norm.

Then the MaxMin distance h_{MM} is defined by:

$$h_{MM} = max\big(d_1, d_2, \ldots d_j \ldots d_n\big), \tag{1}$$

where d_j denotes the minimum distance of the generic spatial unit i from the other units j (with $i \neq j$). As a consequence, the whole territory is connected and there are no isolated spatial units.

The h_{MM} represents the first distance; therefore, it will be called h_{MM}^1.

More formally, $h_{MM}^1 = max(d_1^1, d_2^1, \ldots d_j^1 \ldots d_n^1)$
with $d_j^1 = \min_{i=1\ldots n}(\{d_{ij}^0\}\backslash\{0\})$ with $j = 1 \ldots n$ and $d_{ij}^0 \in \boldsymbol{E^0}$.

Using h_{MM}^1, the generic element ω_{ij}^1 of the first order-spatial weight matrix Ω^1 is determined as follows:

$\omega_{ij}^1 = 1 \; if \; d_{ij}^0 \leq h_{MM}^1$ and $d_{ij}^0 \neq 0$;
$\omega_{ij}^1 = 0$ otherwise
$d_{ij}^0 \in \boldsymbol{E^0}$ and $h_{MM}^1 = max(d_1^1, d_2^1, \ldots d_i^1 \ldots d_n^1)$,
(with $\omega_{ij}^1 \in \Omega^1$).

This first distance h_{MM}^1 is the reference for the Euclidean distance matrix $\boldsymbol{E^1}$ where the generic element d_{ij}^1 is given by:

$d_{ij}^1 = d_{ij}^0 \in \boldsymbol{E^0} \; if \; d_{ij}^0 > h_{MM}^1$;
$d_{ij}^1 = 0$ otherwise.

If we consider the k-distance $h_{MM}^k = max(d_1^k, d_2^k, \ldots d_j^k \ldots d_n^k)$

with $d_j^k = \min_{i=1...n}(\{d_{ij}^{k-1}\}\backslash\{0\})$ with $j = 1...n$ and $d_{ij}^{k-1} \in \boldsymbol{E^{k-1}}$,
for $k = 1...t$,
the generic element ω_{ij}^k of the k-order spatial weight matrix Ω^k is determined as follows:
$\omega_{ij}^k = 1$ *if* $d_{ij}^{k-1} \leq h_{MM}^k$ and $d_{ij}^{k-1} \neq 0$;
$\omega_{ij}^k = 0$ otherwise,
(with $\omega_{ij}^k \in \Omega^k$).

The k-distance h_{MM}^k is the reference for the Euclidean distance matrix $\boldsymbol{E^k}$ where the generic element d_{ij}^k is given by:
$d_{ij}^k = d_{ij}^{k-1} \in \boldsymbol{E^{k-1}}$ *if* $d_{ij}^{k-1} > h_{MM}^k$;
$d_{ij}^k = 0$ otherwise.

As a consequence: $E^0 \supset E^1 \supset ...E^{k-1} \supset E^k \supset ...E^{t-1} \supset E^t$.

So, by iterating this procedure it is possible to obtain the distances from h_{MM}^1 *to* h_{MM}^t.

With h_{MM}^t, all spatial units are linked with each other since the condition that $\forall d_{ij}^{k-1} \leq h_{MM}^t$ $(k = 1...t)$ holds true and the algorithm stops. This distance coincides with the maximum distance in the Euclidean distance matrix $\boldsymbol{E^0}$ $(h_{MM}^t = \max[d_{ij}^0])$. As a consequence, since the condition $d_{ij}^{k-1} > h_{MM}^t$ $(k = 1...t)$ cannot be verified, $\boldsymbol{E^t} = \emptyset$ *and* $\Omega^{t+1} = \emptyset$.

It is important to point out that this procedure generates a "threshold" or "buffer distance" h_{MM}^k (with $k = 1...t$) without imposing any constraint on the number of neighbours; thus, this is not arbitrary but based on the territorial pattern of the spatial units. So, the k-territorial partitions are disjoint from each other, that is, $\Omega^i \cap \Omega^{i+1} = \emptyset$ $(i = 1...t$ with $\Omega^{t+1} = \emptyset)$.

It is important to underline that the algorithm sets the threshold to be the most distant nearest neighbour, consequently each unit has at least one neighbour included so there are no isolated units. For the sake of simplicity and to better illustrate the method, the "real" empirical application refers in fact to the municipality level, we apply this territorial partitioning to the 107 Italian provinces. The MM method applied to the territory creates 22 spatial lags (territorial partitions). The MM distances (h-distance), the relative joints (links) and the names of the provinces that originate the MM distances (province link name) are given in Table 1. Moreover, Figure 1 provides a graphical view of the method.

2.1.2. Partitioning with Constant Step Distance Method

The MM method creates threshold distances in relation to the natural shape of the territory. This criterion can sometimes produce spatial lags that present large gaps in terms of distance. For this reason, we introduce a variant of the MM method with constant increments. We call this procedure the CS method. Suppose we have n spatial units $u_{1......}u_n$ in which we observe $x_{1......}x_n$ data and let $\boldsymbol{E^0}$ be a $n \times n$ matrix of the Euclidean distances between these units such that $d_{ij}^0 = \|u_i - u_j\|_2$ (with $d_{ij}^0 \in E^0$, $i = 1...nj = 1...n$), where $\|\cdot\|_2$ is the Euclidean norm.

As in the previous procedure we determine the first distance of MM which in this method coincides with the first distance $\left(h_{cs}^1 \equiv h_{MM}^1\right)$. If we set the increment equal to the first distance, we can write the following relation for the generic spatial lag (k):

$$\mathrm{h}_{\mathrm{cs}}^{\mathrm{k}} = \mathrm{kh}_{\mathrm{cs}}^{1} \text{ (with } k = 1...t). \tag{2}$$

Table 1. Details of the MM method applied to the 107 Italian provinces[1].

Spatial Lag	h-distance (Km)	Links	Name of the provinces that determine the h-distance
1	77.96	516	(‘Palermo’, ‘Trapani’)
2	116.74	554	(‘Palermo’, ‘Catania’)
3	375.75	4140	(‘Grosseto’, ‘Oristano’)
4	417.92	596	(‘Palermo’, ‘Cagliari’)
5	454.58	510	(‘Chieti’, ‘Messina’)
6	492.57	518	(‘Latina’, ‘Cremona’)
7	528.30	452	(‘Pistoia’, ‘Sud Sardegna’)
8	570.48	528	(‘Crotone’, ‘Perugia’)
9	618.61	522	(‘Salerno’, ‘La Spezia’)
10	666.72	490	(‘Forli’-Cesena’, ‘Lecce’)
11	712.72	424	(‘Grosseto’, ‘Siracusa’)
12	759.24	342	(‘Salerno’, ‘Asti’)
13	804.87	302	(‘Pisa’, ‘Siracusa’)
14	850.90	292	(‘Ravenna’, ‘Siracusa’)
15	899.84	280	(‘Modena’, ‘Siracusa’)
16	942.64	268	(‘Parma’, ‘Siracusa’)
17	985.85	238	(‘Verona’, ‘Siracusa’)
18	1036.66	222	(‘Brescia’, ‘Siracusa’)
19	1087.97	102	(‘Lecco’, ‘Siracusa’)
20	1111.11	34	(‘Bolzano’, ‘Siracusa’)
21	1153.71	10	(‘Aosta’, ‘Siracusa’)
22	1154.03	2	(‘Aosta’, ‘Ragusa’)

[1] It is important to clearly explain that the MM distance method creates distances in relation to the spatial configuration of territorial units. This criterion can sometimes produce spatial lags that present large gaps in terms of distance. If we look, for example, to Table 1 a big jump is evident between lags 2 and lags 3 due to the link created between the Sardinia region (that is an island) and the rest of Italy. For a more technical explanation, please see Appendix (Section 6).

The procedure stops when $h^t_{cs} \geq \max\,[d^0_{ij}]$ where $d^0_{ij} \in \boldsymbol{E^0}$

In the h^t_{cs} distance all the spatial units are linked with each other. For this method the details and a graphical illustration are also provided considering its application to the 107 provinces of Italy (Table 2 and Figure 2).

2.2. *From the Territorial Partitions to the Spatial Gini Index (SGI)*

The properties of the territorial partitions of the two distance methods discussed above makes the procedure compatible with the structure of the Gini index according to the definition of the ratio of the areas (Mucciardi and Bertuccelli 2007; Mucciardi 2008b). Indeed, the Gini index can then be thought of as the ratio of the area that lies between the line of equality and the Lorenz curve over the total area under the line of equality. Following the same approach, SGI can therefore be considered as the ratio between the area of spatial autocorrelation on the total area of the square of side 1.

We define $J_{(k)}$, the cumulated percentage of the total connectivity of the units in the generic distance h^k_{MM}, as

$$J_{(k)} = \frac{\sum_i^n \sum_j^n \omega_{ij}^k}{A} \downarrow \text{ with } k = 1 \ldots t \quad (J_{(0)} = 0 \text{ and } J_{(t)} = 1) \qquad (3)$$

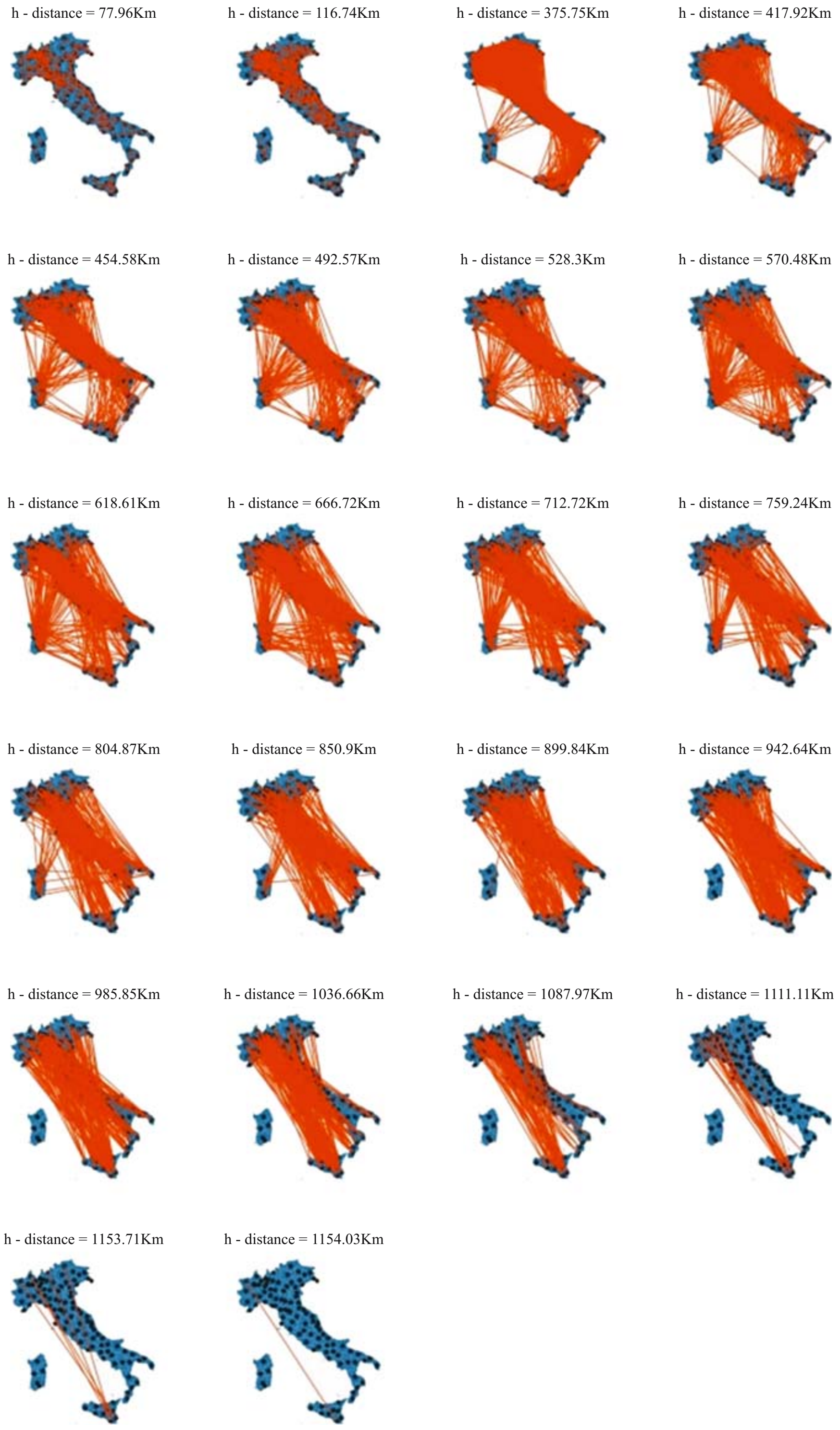


Fig. 1. Graphical view of the MM method applied to the 107 Italian provinces.

Table 2. Details of the CS distance method applied to the 107 provinces of Italy.

Spatial Lag	h-distance (Km)	Links
1	77.96	516
2	155.92	1170
3	233.89	1334
4	311.85	1250
5	389.81	1142
6	467.77	1066
7	545.74	1028
8	623.70	884
9	701.66	746
10	779.62	596
11	857.58	492
12	935.55	478
13	1013.51	412
14	1091.47	184
15	1169.43	44

where ω^k_{ij} denote the interconnection links (that is, the contiguous spatial units), in the generic k-territorial partition h^k_{MM} and $A = \Sigma^t_{i=1}\Omega^i = \Sigma^t_{k=1}\Sigma^n_{i=1}\Sigma^n_{j=1}\omega^k_{ij} = n(n-1)$ is the maximum number of links it is possible to obtain from a particular territorial configuration Please note that in the equations the arrows symbol means that the percentage are cumulated. Furthermore, we recall that the connections of the spatial units with themselves are excluded from the maximum number of links.

In the same way, we define $V_{(k)}$, the cumulated percentage of the variability of the phenomenon X "absorbed" by the linked elements in the distance h^k_{MM}, as

$$V_{(k)} = \frac{\Sigma^n_i \Sigma^n_j \left(x_i - x_j\right)^2 \omega^k_{ij}}{D} \downarrow \text{ with } k = 1 \ldots t \quad \left(V_{(0)} = 0 \text{ and } V_{(t)} = 1\right) \qquad (4)$$

where ω^k_{ij} are as before and

$$D = \Sigma^t_{k=1}\Sigma^n_{i=1}\Sigma^n_{j=1}\left(x_i - x_j\right)^2 \omega^k_{ij}.$$

Referring to a territorial system comprising n spatial units in which we observe $x_1 \ldots x_n$ data, the Spatial Gini Index (SGI) will be defined as

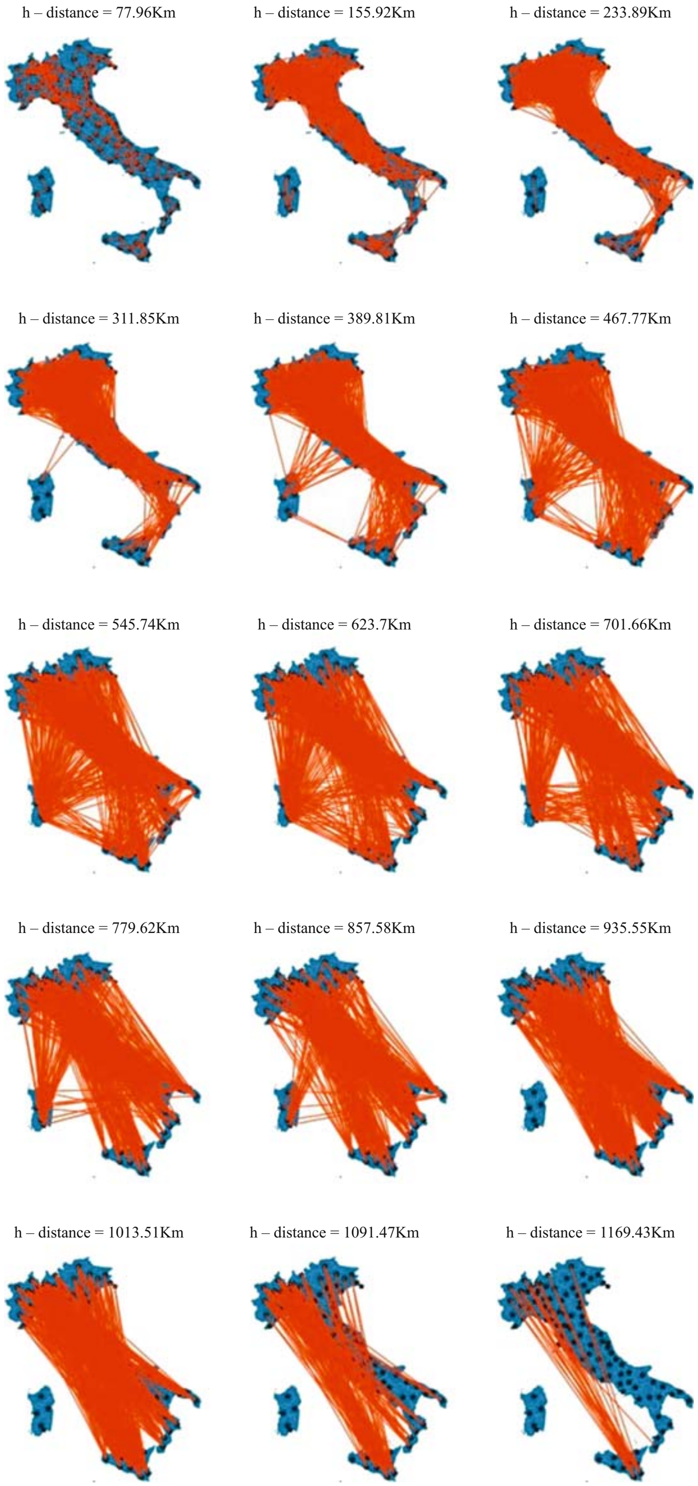
$$SGI = 1 - 0.5 \sum_{k=1}^{t} \left(V_{(k)} + V_{(k-1)}\right)\left(J_{(k)} - J_{(k-1)}\right)$$

with

$$V_{(k)} = \frac{\Sigma^n_i \Sigma^n_j \left(x_i - x_j\right)^2 \omega^k_{ij}}{D} \downarrow \quad \text{and} \quad J_{(k)} = \frac{\Sigma^n_i \Sigma^n_j \omega^k_{ij}}{A} \quad k = 1 \ldots t. \qquad (5)$$

Its construction is based on the computation of the area of spatial autocorrelation, which has been proposed, to the best of our knowledge, (see for more details, Alleva 1987; Mucciardi and Bertuccelli 2007; Mucciardi 2008b).

Following the Lorenz curve, which is the basis of the Gini index, in a condition of no spatial autocorrelation, the cumulated percentage of variability $V_{(k)}$ should not differ from the cumulated percentage of connectivity $J_{(k)}$. SGI can assume the minimum value of 0



h – distance = 77.96Km

h – distance = 155.92Km

h – distance = 233.89Km

h – distance = 311.85Km

h – distance = 389.81Km

h – distance = 467.77Km

h – distance = 545.74Km

h – distance = 623.7Km

h – distance = 701.66Km

h – distance = 779.62Km

h – distance = 857.58Km

h – distance = 935.55Km

h – distance = 1013.51Km

h – distance = 1091.47Km

h – distance = 1169.43Km

Fig. 2. Graphical illustration of the CS distance method applied to the 107 provinces of Italy.

and the maximum value of 1, but in terms of spatial autocorrelation we distinguish three cases:

1. Case of "negative spatial autocorrelation": if the relative contributions in terms of variability are larger than the contribution in terms of connectivity in the generic distance h_{MM}^{k}, we will graphically obtain a "convex curve" with respect to the ordinate axis and $0 \leq \mathrm{SGI} < 0.5$,
2. Case of "no spatial autocorrelation": if the relative contributions in terms of variability and connectivity increase proportionally to the variation of distance h_{MM}^{k}, we will graphically obtain a straight line at a perfect angle of 45° and $\mathrm{SGI} = 0.5$ (area of the curve exactly equal to 0.5), and
3. Case of "positive spatial autocorrelation": if the relative contributions in terms of variability are smaller than the contribution in terms of connectivity in the generic distance h_{MM}^{k} we will graphically obtain a "concave curve" with respect to the ordinate axis and $0.5 < \mathrm{SGI} \leq 1$.

To better understand the relation between the concept of a Lorenz curve and that of spatial correlation, Figure 3 shows three scenarios of spatial autocorrelation with the related range of values for SGI.

From a geometric point of view, these three forms of spatial autocorrelation may be assessed, as the h_{MM}^{k} (h-distance) varies, by considering the tangent of the angle formed by the straight line with the x-axis:

$$\tan^{k}(\alpha) = \frac{V_{(k)}}{J_{(k)}} \quad k = 1 \ldots t. \tag{6}$$

So, we can have:

1. $\tan^{\mathrm{k}}(\alpha) < 1$ (angle $< 45°$) indicating positive spatial autocorrelation,
2. $\tan^{\mathrm{k}}(\alpha) = 1$ (angle $= 45°$) indicating no spatial autocorrelation, and
3. $\tan^{\mathrm{k}}(\alpha) > 1$ (angle $> 45°$) indicating negative spatial autocorrelation.

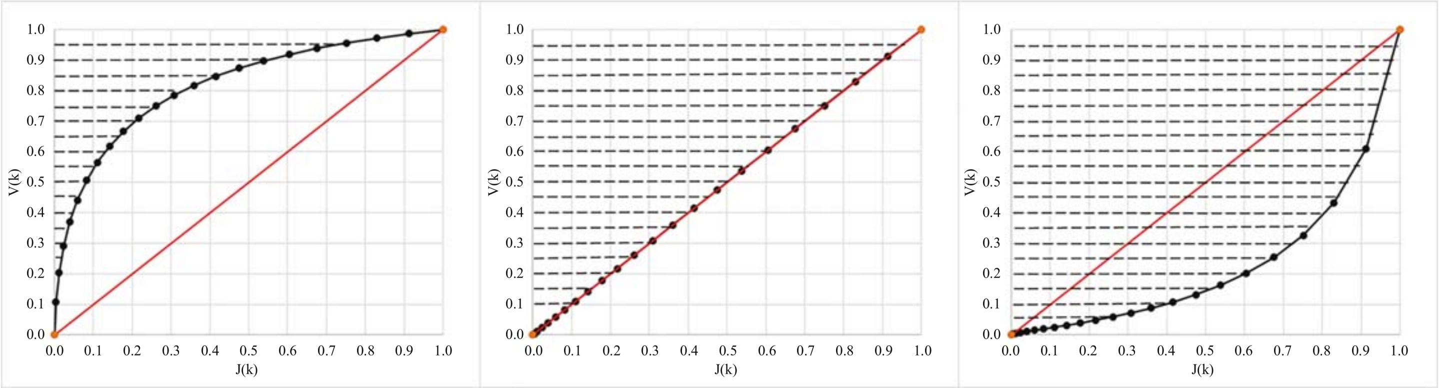
For the calculation of the angle, we use the arctan function. We recall that the arctan function is the inverse of the tangent function. It returns the angle whose tangent is a given number. Figure 4 shows the expected trend of the arctan function (degrees) as the h_{MM}^{k} vary for the three scenarios of spatial autocorrelation.

By using the CS distance method instead of MM in all these relations, we obtain the SGI with this territory partition method. In the context of this research, we will distinguish the calculation of the SGI using this notation:

1. SGI_{MM} when the k-partitioning method is that of the MM distance, and
2. SGI_{CS} when the k-partitioning method is that of the CS distance

3. The Decomposition Method of Rey and Smith in the Framework of SGI

A spatial decomposition of the Gini coefficient has recently been proposed by Rey and Smith (2013). In this work the authors suggest an alternative approach towards considering the joint effects of inequality and spatial autocorrelation that relies on a decomposition of the classic Gini coefficient. This decomposition involves the splitting of



(A) Simulated data for 22 spatial lags with three assumed scenarios: SGI = 0.2 (case of negative spatial autocorrelation); SGI = 0.5 (case of no spatial autocorrelation); SGI = 0.8 (case of positive spatial autocorrelation)

Fig. 3. *Curve of spatial autocorrelation (dashed line): 1) case of negative spatial autocorrelation – 0 < SGI <0.5 (left); 2) case of no spatial autocorrelation – SGI = 0.5 (centre); 3) case of positive spatial autocorrelation 0.5 < SGI <1 (right) (A).*

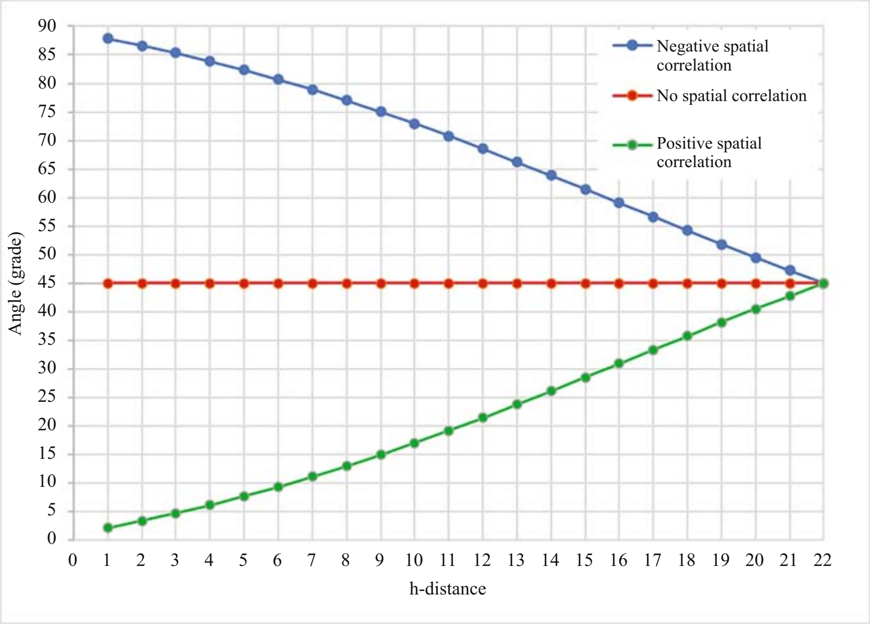


Fig. 4. Expected trend of the arctan function (degrees) as the h-distances vary for the three scenarios of spatial autocorrelation – (simulated data).

the Gini index into two mutually exclusive components: into contiguous units (w_{ij}) and the non-contiguous units ($1 - w_{ij}$).

Equation 7 shows this decomposition.

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{2n^2\mu} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\left|x_i - x_j\right|}{2n^2\mu} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(1 - w_{ij})\left|x_i - x_j\right|}{2n^2\mu}. \quad (7)$$

According to SGI we extend this decomposition applying to the two distance methods (MM and CS) shown above. Therefore, it is possible to make this decomposition of each "*k-territorial partition*" imposed by the h^k-distances.

Now, if we denote by G_C^k the Gini index calculated inside the contiguous units in the h^k -distances and by G_{NC}^k the Gini index calculated in the non-contiguous units (or outside the contiguous units) in the h^k-distances, we can rewrite the Gini index with the Rey and Smith decomposition (G_T^k):

$$G_T^k = G_C^k + G_{NC}^k \quad (k = 1 \ldots t). \quad (8)$$

As a consequence,

$$G_T^k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\mu} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij}^k|x_i - x_j|}{2n^2\mu} + \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(1 - \omega_{ij}^k)|x_i - x_j|}{2n^2\mu} \quad (9)$$

$$(k = 1 \ldots t)$$

with

$$G_C^k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\omega_{ij}^k\left|x_i - x_j\right|}{2n^2\mu};\ G_{NC}^k = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(1-\omega_{ij}^k)\left|x_i - x_j\right|}{2n^2\mu}$$

and $G_T^k \equiv G\ (\forall k = 1\ldots t)$.

To distinguish the decomposition based on the partition used (MM and CS), we will differentiate the calculation of the G_T^k index using this notation:

1. $G_{T(MM)}^k = G_{C(MM)}^k + G_{NC(MM)}^k$ when the *k*-partitioning method is that of the MM distance
2. $G_{T(CS)}^k = G_{C(CS)}^k + G_{NC(CS)}^k$ when the *k*-partitioning method is that of the CS distance.

Our task here is both to propose an extension of the decomposition method of Rey and Smith in the framework of the two distance approaches (MM and CS) here discussed and to evaluate similarities and differences in the achieved results. Our general idea is that all these measures can be used in a complementary way in measuring the spatial concentration of the population. This is somewhat similar to what happens with measuring residential segregation, where the use of different measures and approaches is highly recommended (Brown and Chung 2006)

4. Empirical Application

In this section we show the results of the empirical application. The application regards three points in time, 2002, 2010 and 2018, and it is realized at the municipality level, the finest territorial scale possible, that is to say almost 7,900 spatial units. The application was done for two groups of population selected on the basis of country of citizenship: Italians and foreigners. As is known, these two populations typically have different geographical patterns of spatial distribution and therefore they are particularly useful for our tasks (Massey and Denton 1988). The data are provided by the Italian National Institute of Statistics (Istat) and disseminated by the institutional website. In more detail, we use data on the resident population broken down by the country of citizenship (Italian/Foreign) for the years 2002, 2010, and 2018.

4.1. The Spatial Gini Index

4.1.1. The MM Approach

In Figure 5 we can appreciate the evolution over time of the level of the classical Gini index and the spatial Lorenz curve of Italian and foreign population in the observed period. In the first year, 2002, the foreign population in Italy was about 1.3 million and the level of its spatial concentration was lower than the Italian population (SGI = 0.446 versus SGI = 0.485). In the following years the Italian population remained quite stable and so did its level of spatial concentration: SGI = 0.485 in 2002 and SGI = 0.483 in 2018. In contrast, the foreign population grew significantly, reaching more than 5 million in 2018. Its level of spatial concentration grew too: SGI = 0.446 in 2002 and SGI = 0.450 in 2018. The dynamics of the arctan functions (Figure 6) inform us about the evolution of the level of spatial autocorrelation for both populations in the selected years for each *h*-distance. As can

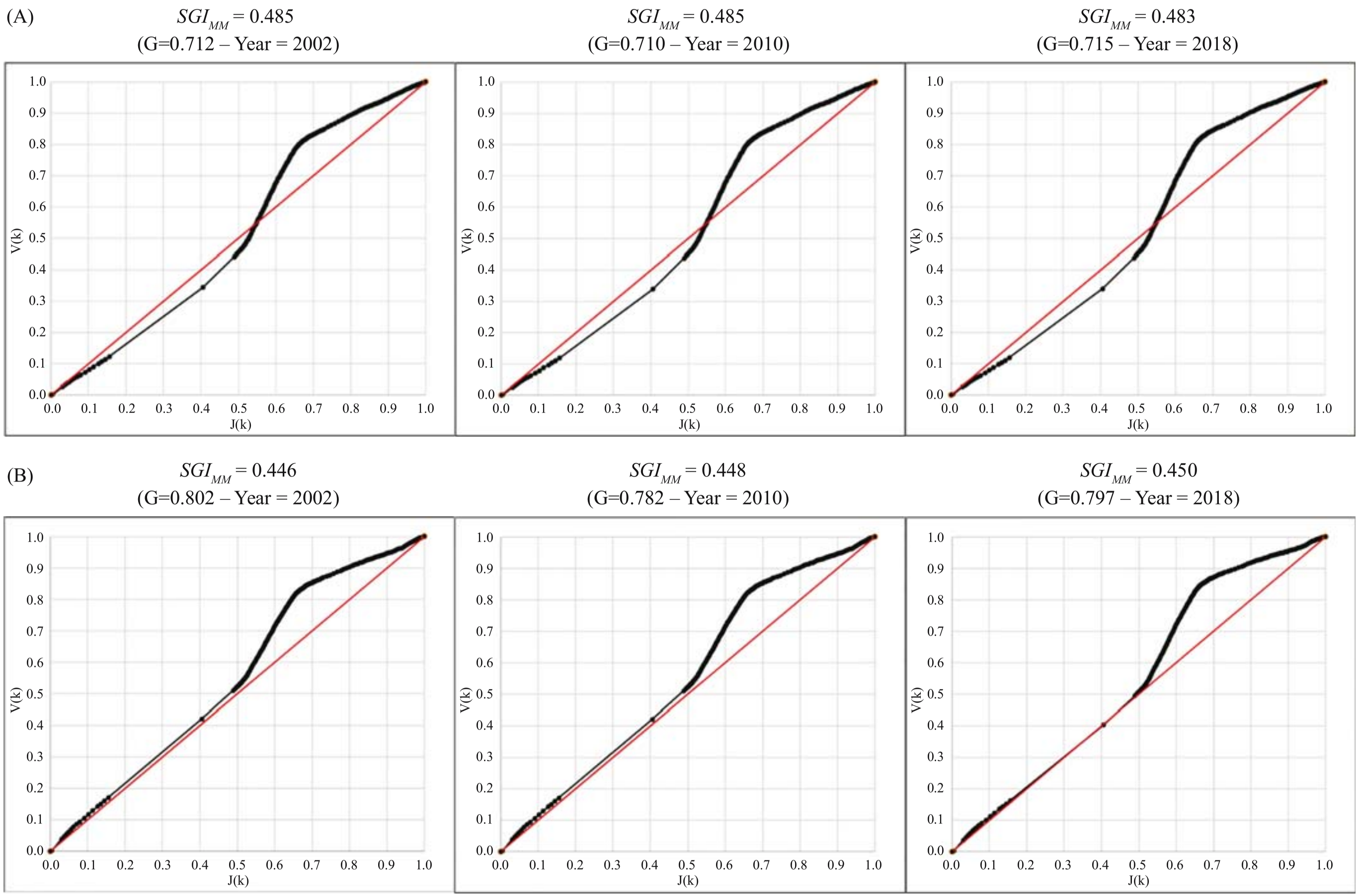


Fig. 5. Spatial Lorenz Curve, SGI_{MM} and aspatial Gini's G (in brackets). Italian population (A) and foreign population (B).

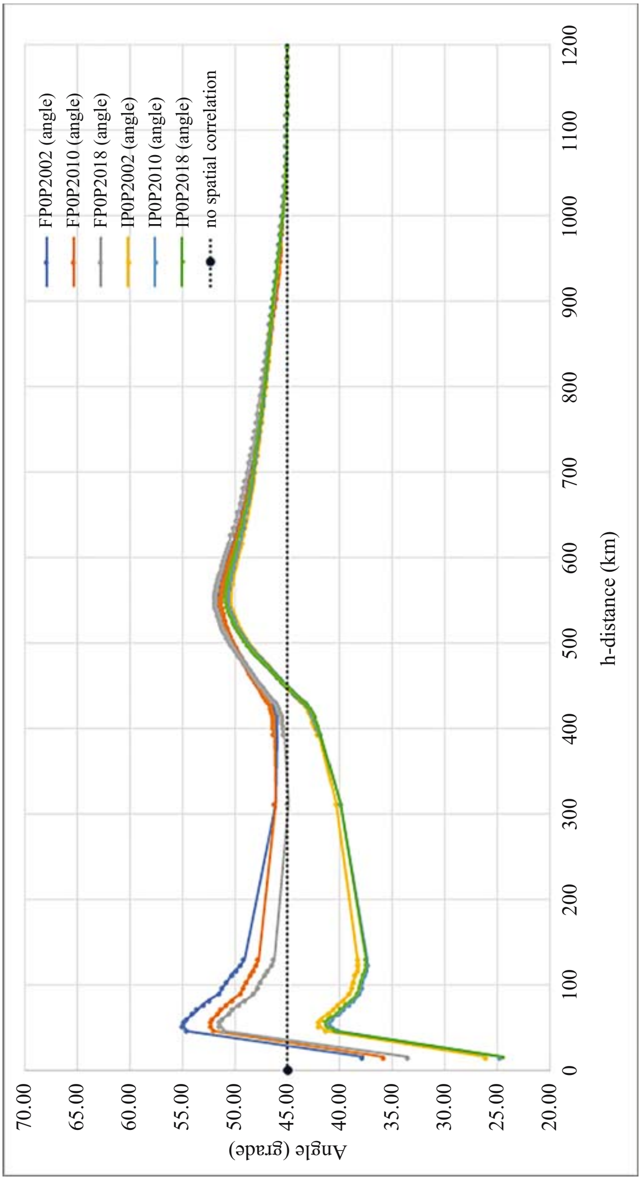


Fig. 6. Value of arctan (degrees). Italian (IPOP) and foreign (FPOP) population (MM Distance Method).

be clearly seen, the two populations start (16 km) from a different condition of positive spatial autocorrelation: greater positive spatial autocorrelation for the Italian population and less positive spatial autocorrelation for the foreign population (see the angles less than 45 degrees in Figure 6). From the second *h*-distance (47 km) the situation changes: the positive spatial autocorrelation effect is attenuated for the Italian population while the foreign population shows spatial instability with negative values of arctan (see in Figure 6 the angles greater than 45 degrees for foreign populations only). By increasing the *h*-distance, a process of convergence towards a level of no spatial autocorrelation comes up. This condition is obtained starting from an *h*-distance equal to 400 km and obtained when the *h*-distance is equal to 1,000 km. From the point of view of the classic Gini index (*G*), we note an inverse behaviour: lower values of SGI are followed by higher values of *G*. Therefore, from these results it emerges that the foreign population, although concentrated at an "aspatial level" (*G*), instead shows a moderate tendency to have a negative autocorrelation with the territory. The Italian population shows a lower level of the aspatial concentration (*G*). This behaviour manifests itself in the SGI with a tendency towards no spatial autocorrelation (SGI tends to 0.5). However, we verify the spatial behaviour of these two populations by calculating the Moran's *I* index for the same *h*-distances (Figure 7). As we can see at 16 km (first *h*-distance) the different manifestation of positive spatial autocorrelation is confirmed, albeit in a more attenuated manner (Moran's *I* indices, although statistically significant, are very small). The spatial autocorrelation level then converges to the expected values exactly from a distance of 400 km.

4.1.2. Constant Step Approach

From a territorial partitioning point of view, the main difference between the two methods (MM and CS) is in the number of *h*-distances produced: 162 for MM and 76 for CS (see Appendix, Subsection 6.1, for more details). However, despite this difference in the number of *h*-distances, the results of the CS procedure are quite similar to the ones obtained by MM (see Subsection 4.1.1) with slight substantial differences. Therefore, the explanation of the spatial concentration process of the Italian and foreign populations in the years considered (2002, 2010 and 2018) remains almost unchanged with respect to the SGI index calculated with the MM method (see Figure 8–10).

In general, and based on some empirical evidence known in the literature (Miller 2004), positive autocorrelations on a small spatial scale and negative autocorrelations on a larger spatial scale are all in all expected results. The first law of geography states that "everything is related to everything else, but near thing is more related than distant things" (Tobler 1970, 235). As known, scale can play important effect on results, nevertheless our results seem to be coherent with Tobler's law. Indeed, we know that the spatial distribution of human population, is a process particularly affected by such kind of "spatial" effects (distance and scale). This is particularly true for the foreign population that tend to have specific settlement models to maximize opportunity and minimize cost. As know the scientific debate about that is wide and rich but, at least for Italy, there are clear evidence about the attractive role played by urban areas on foreign and immigrant population (Strozza et al. 2016) and on the spatial heterogeneity and dependence that characterized the settlement models of foreign communities at different spatial scales (Benassi et al. 2019, 2020).

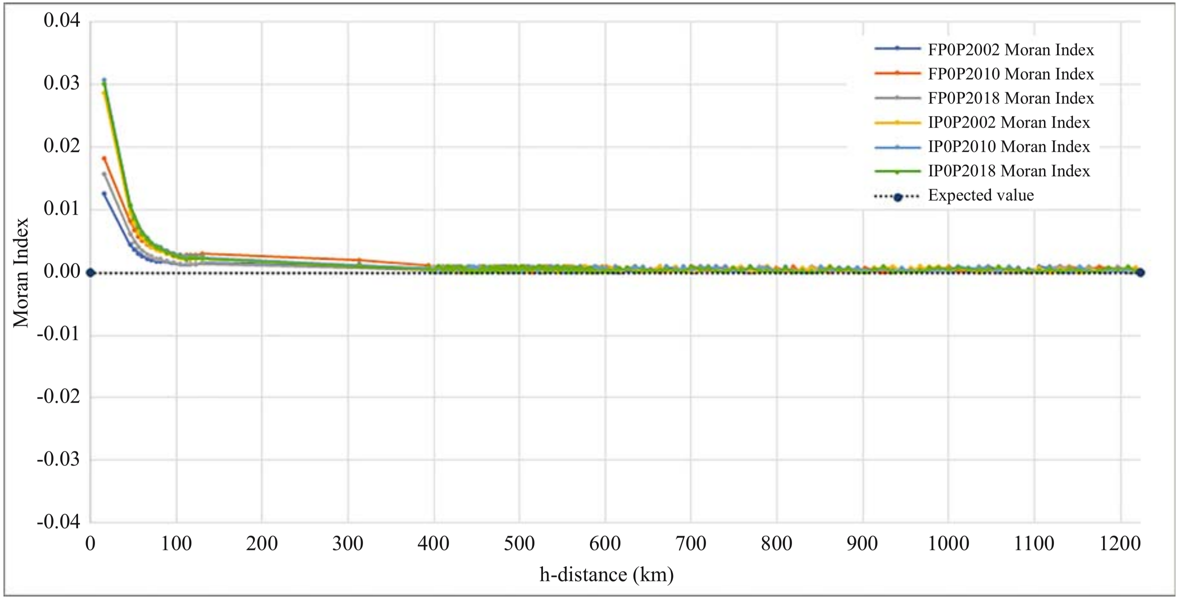


Fig. 7. Moran's I index. Italian (IPOP) and Foreign (FPOP) populations (a) (MM Distance Method).
(a) Moran's I indices are almost all statistically significant ($p < 0.05$). All the Moran's I indices are processed with Arcgis10.8 software for the same MM distances used for SGI. Data are available upon request.

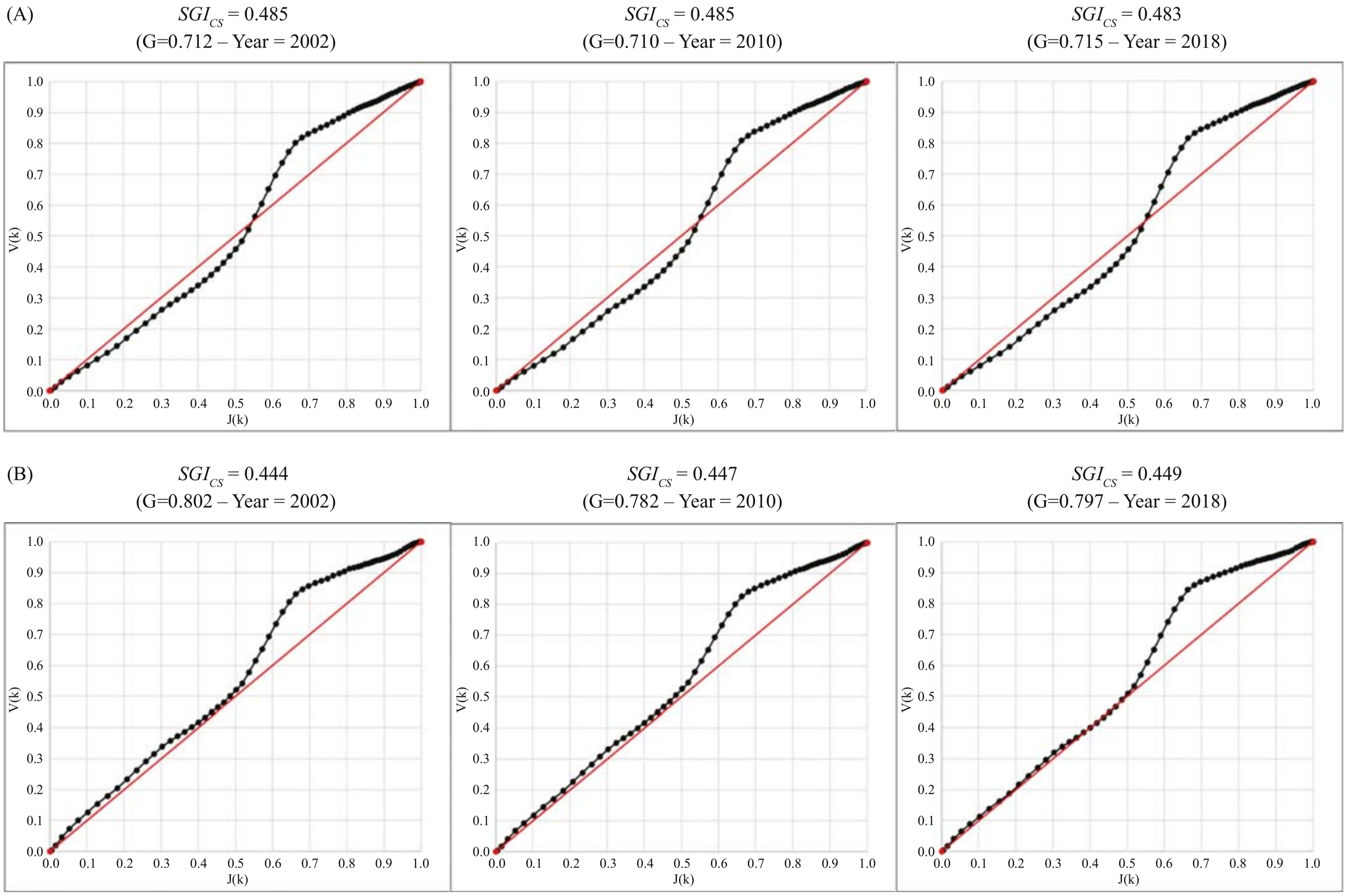


Fig. 8. Spatial Lorenz Curve, SGI_{CS} and aspatial Gini's G index (in brackets). Italian population (A) and foreign population (B).

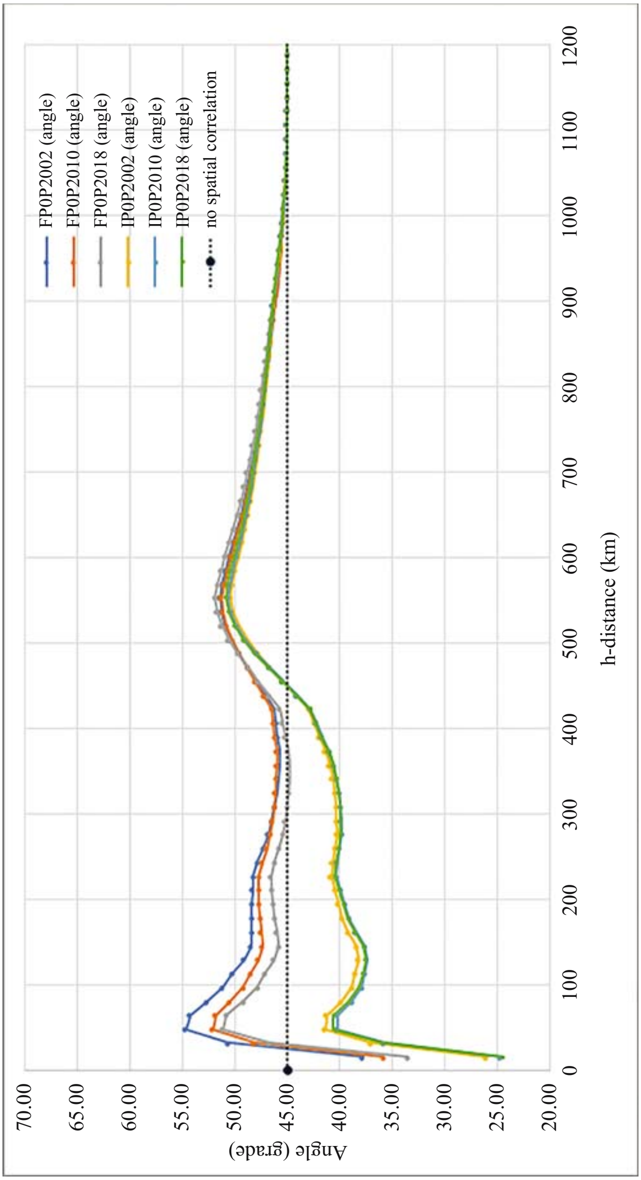


Fig. 9. Values of arctan (degrees). Italian (IPOP) and foreign (FPOP) population (CS Distance Method).

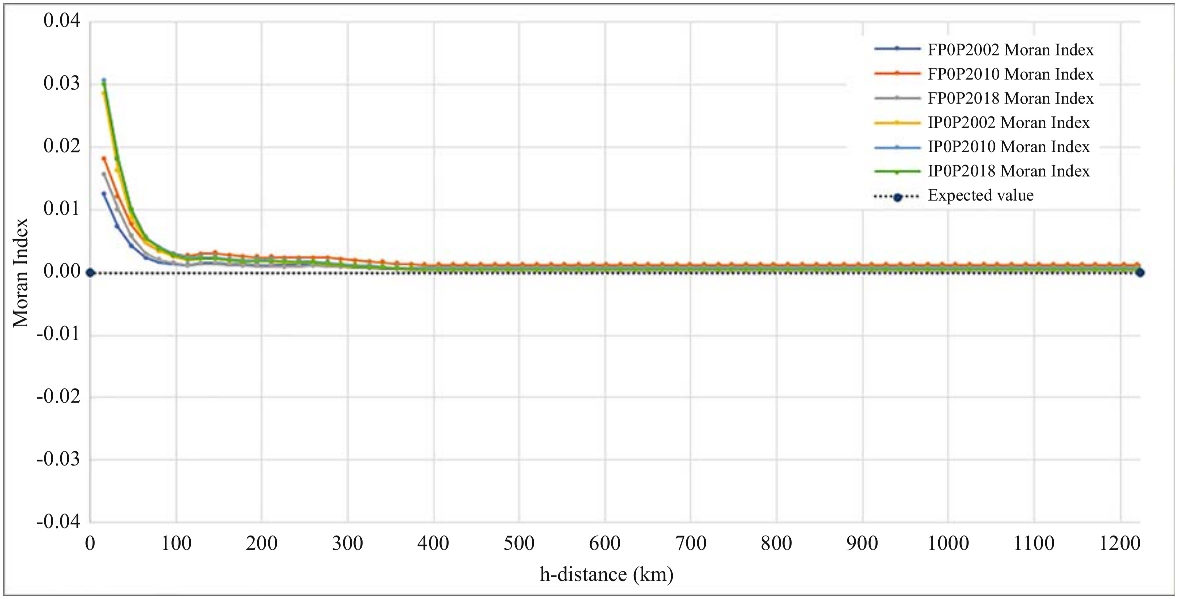


Fig. 10. Moran's I index. Italian (IPOP) and Foreign (FPOP) population (a) (CS Distance Method).

(a) Moran's I indices are almost all statistically significant ($p < 0.05$). All the Moran's I indices were processed with Arcgis10.8 software for the same CS distances used for SGI. Data are available upon request.

4.2. Rey and Smith Decomposition Method With MM and CS Approaches

As mentioned above, let's extend the Rey and Smith decomposition (G_T^k) applying the two distance methods: $G_{T(MM)}^k$ with MM approach and $G_{T(CS)}^k$ with CS approach. As can be seen from Figures 11 and 12, considering the same h-distances with which we build SGI, the G_T^k in the spatial version is less sensitive to the quantification of inequality in the various spatial lags. In fact, the components of the Gini index (in the contiguous G_C^k and non-contiguous G_{NC}^k units) remain almost unchanged (the lines are almost parallel). Only considering the $G_{T(MM)}^k$, around 300 km, we have a peak of the Gini index in the component calculated in the contiguous units (G_C^k). In our opinion this occurs because the two indices measure spatial inequality differently. While SGI is a global measure of spatial inequality, that is, it is (completely) evaluated considering all the spatial lags (k), the G_T^k (with $G_T^k = G_C^k + G_{NC}^k$) is evaluated in each k-spatial lag. So, from this point of view, SGI and the G_T^k can be seen as complementary measures (or even tools) to better grasp and detect the level of (spatial) concentration and its dynamics for a given population.

5. Discussion and Conclusions

Measuring the concentration of population is an old and traditional activity of applied statistics. Many developments have been proposed from the seminal contributions of Lorenz and Gini. In relative recent times, also due to the GIS revolution, more attention has been paid to the spatial aspects of concentration, especially when it refers to a human population (Arbia 2001). The present article lies in this stream of the literature, proposing a new approach for measuring the spatial concentration of a human population, based on a spatial version of Gini's G index, which we call the Spatial Gini Index (SGI). In any kind of spatial approach, the definition of spatial neighbourhood and spatial weight matrixes are crucial and pivotal issues. This article proposes two new approaches to partitioning the territory, and therefore obtains two different kinds of spatial connectivity: one is based on a MM approach and allows computing the $\mathrm{SGI_{Mm}}$ version of the SGI. The other is based on a CS approach, and it is the basis for the computation of $\mathrm{SGI_{CS}}$. From the results obtained in the previous sections, SGI results are different from the ones produced by classic (aspatial) Gini's G index, indicating: (1) the importance of the spatial dimension in detecting the concentration of a population in space; (2) higher values of the level of concentration when measured using a traditional aspatial approach; (3) that this difference (G vs SGI) is higher for the foreign population. The higher values reached by the aspatial version of the concentration index are due to the fact that this approach is essentially based on the statistical concept of variability separated from the influence of the territory. So, this kind of measuring of (aspatial) concentration assumes that the space is independent of the distribution of the variable (and vice versa). From this point of view, the SGI index can be viewed as, at least, a complementary tool for better measuring and detecting the spatial concentration of a population. In particular, given the peculiarity of the spatial distribution of the foreign population in a destination country like for example Italy (Strozza et al. 2016) and the relevance of the spatial concentration for such a population (Reardon and O'Sullivan 2004), the proposed approach seems to add new methodological perspectives on measuring spatial concentration. In evaluating SGI, we also used well known spatial statistical measures of global spatial autocorrelation (Moran's I). The index has been

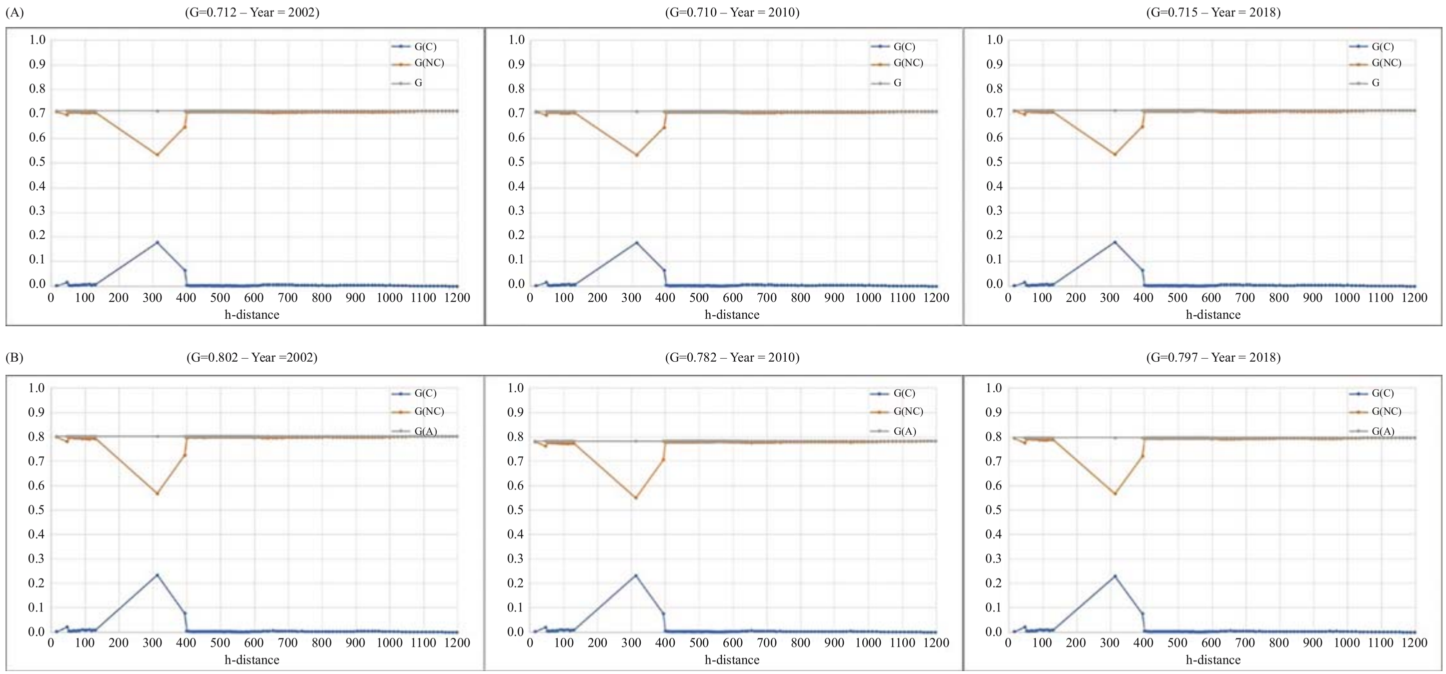


Fig. 11. $G^{k}_{T(MM)}$ *and aspatial Gini's G index (in brackets). Blue line –* $G^{k}_{C(MM)}$*; Orange line* $G^{k}_{NC(MM)}$*; Grey line – G (Italian population (A) and foreign population (B)).*

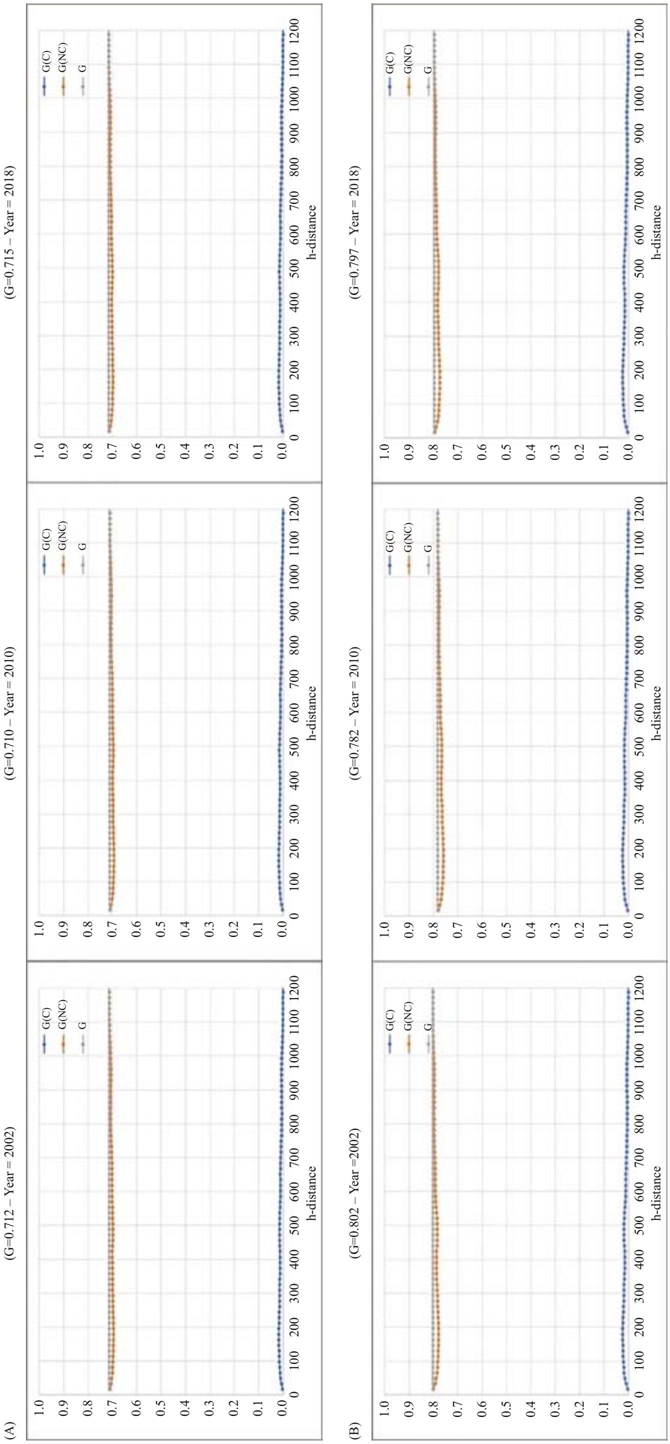


Fig. 12. $G^k_{T(CS)}$ and aspatial Gini's G index (in brackets). Blue line – $G^k_{C(CS)}$; Orange line $G^k_{NC(CS)}$; Grey line – G (Italian (A) and foreign population (B)).

computed on the resident population (Italian and foreigners) of Italian municipalities in 2002, 2010 and 2018. The results indicate that, at least using population as a variable, the level of global spatial autocorrelation is quite low and obviously tends to decrease with increasing *h*-distance. Another point faced in this article is the extension of the decomposition method of the Gini's *G* index recently proposed by Rey and Smith (2013) in the context of the territorial partitioning. Here the main objective was to propose an extension of this recent method in the framework of the MM and CS distances procedure and to evaluate similarities and differences between them both. Future developments should address the extension/optimization of the Monte Carlo test (so far only tested for 100 statistical units) and solving problems of computational requirements. There is also the possibility of using "classical binary weight" based on portions of shared boundaries of spatial unity (e.g., Queen and Rook method) and "non binary weights" in the spatial matrix (e.g., a kernel matrix with a distance decay function, etc.) and do simulation experiments in order to grasp the behaviour of SGI in terms of statistical distribution. In this direction it will be possible to make comparisons between the different methods of partitioning of the territory. Further developments can regard the local decomposition of the SGI and proposing other functionalities like semivariogram and similar. From an interpretative point of view, we have to underline that high levels of spatial concentration of foreign population can lead to different processes and behaviours that can act as detrimental to social cohesion. In conclusion, researching new approaches to measuring the spatial concentration of the human population is still an open challenge and improvements are currently in progress by the authors.

6. Appendix

6.1. Details of the SGI Procedure

6.1.1. MM Distance

Number of municipalities = 7,890 – Total links = 62,244,210

Spatial Lag	h-distance (km)	Links	% Links	J(k)	Name of the Italian municipalities that determine the h-distance
1	16.3	269852	0.004	0.004	(Campagnatico, Grosseto)
2	47.0	1609960	0.026	0.030	(Piraino, Malfa)
3	52.7	416318	0.007	0.037	(Follonica, Marciana)
4	57.0	337466	0.005	0.042	(Bibbona, Porto Azzurro)
5	61.7	373534	0.006	0.048	(San Severo, Vieste)
6	67.3	470188	0.008	0.056	(Campo nellElba, Monteverdi Marittimo)
7	72.0	417136	0.007	0.063	(Statte, Nard)
8	77.0	447572	0.007	0.070	(Martina Franca, Surbo)
9	83.5	600656	0.010	0.079	(Massafra, Carmiano)
10	91.4	766296	0.012	0.092	(Alberobello, Lecce)
11	98.3	675754	0.011	0.103	(Cavallino, Massafra)
12	105.6	727020	0.012	0.114	(Alberobello, Vernole)

Continued

Spatial Lag	h-distance (km)	Links	% Links	J(k)	Name of the Italian municipalities that determine the h-distance
13	113.9	852748	0.014	0.128	(Alberobello, Melendugno)
14	119.0	521240	0.008	0.136	(Massafra, Cannole)
15	124.5	566798	0.009	0.145	(Massafra, Otranto)
16	131.2	682664	0.011	0.156	(Ortueri, Palau)
17	312.8	15407126	0.248	0.404	("Campo nellElba", Ghilarza)
18	394.9	5325726	0.086	0.489	(Erice, Buggerru)
19	400.9	375570	0.006	0.496	(Castellammare del Golfo, Iglesias)
20	405.3	278344	0.004	0.500	(Mazara del Vallo, Portoscuso)
21	407.7	152894	0.002	0.502	(Santa Ninfa, San Giovanni Suergiu)
22	411.7	252966	0.004	0.507	(Vita, Calasetta)
23	413.8	135790	0.002	0.509	(Giungano, Portopalo di Capo Passero)
24	416.8	195508	0.003	0.512	(Santa Ninfa, Calasetta)
25	419.5	171870	0.003	0.515	(Cagli, Vignone)
26	421.9	160844	0.003	0.517	(Partanna, Calasetta)
27	424.2	146760	0.002	0.520	(Torrita di Siena, Perito)
28	427.4	209442	0.003	0.523	(Umbertide, San Bernardino Verbano)
29	429.8	160176	0.003	0.525	("Campo nellElba", Masainas)
30	431.9	132634	0.002	0.528	(Camerano, Taceno)
31	434.1	146030	0.002	0.530	(Orbetello, Sala Consilina)
32	436.5	160766	0.003	0.533	(Borgo Pace, "SantAngelo a Fasanella")
33	438.9	155548	0.002	0.535	(Guanzate, Montecassiano)
34	440.8	129086	0.002	0.537	(Borgo Pace, Corleto Monforte)
35	443.1	154146	0.002	0.540	(Teora, Ragusa)
36	446.0	191916	0.003	0.543	(Città di Castello, Futani)
37	448.9	197168	0.003	0.546	(Castiglione del Lago, Domodossola)
38	451.4	165464	0.003	0.549	(Bibbona, Castelfranci)
39	453.5	143382	0.002	0.551	(Castiglione del Lago, Bognanco)
40	455.5	136260	0.002	0.553	(Senigallia, Nemoli)
41	457.3	116818	0.002	0.555	(Varese, Viterbo)
42	459.9	183474	0.003	0.558	(Cagli, Trasquera)
43	462.1	144276	0.002	0.560	(Città di Castello, San Giovanni a Piro)
44	464.0	136254	0.002	0.562	(Peglio, Grassano)
45	466.0	137450	0.002	0.565	(Rancio Valcuvia, Sirolo)
46	467.8	122780	0.002	0.567	(Cassano Valcuvia, Sirolo)
47	469.8	140424	0.002	0.569	(Villongo, Montefino)
48	471.5	115082	0.002	0.571	(Treviglio, Mandela)
49	473.9	167052	0.003	0.573	(Colli al Metauro, Castelsaraceno)
50	476.2	160610	0.003	0.576	(Pieve Santo Stefano, Irsina)
51	478.3	148072	0.002	0.578	(Tavullia, Casaletto Spartano)

Continued

Spatial Lag	h-distance (km)	Links	% Links	J(k)	Name of the Italian municipalities that determine the h-distance
52	480.3	141082	0.002	0.581	(San Godenzo, Ricigliano)
53	482.6	164560	0.003	0.583	(Orbetello, Latronico)
54	485.1	172450	0.003	0.586	(San Giovanni Valdarno, Genzano di Lucania)
55	487.8	197618	0.003	0.589	("Colle di Val dElsa", "SantAngelo Le Fratte")
56	489.7	131192	0.002	0.591	(Castiglione della Pescaia, Anzi)
57	491.6	135422	0.002	0.593	(Costa Serina, Montefino)
58	493.4	126432	0.002	0.595	(Cappelle sul Tavo, Librizzi)
59	495.3	134642	0.002	0.598	(Monterotondo Marittimo, Brienza)
60	497.8	171850	0.003	0.600	(Cesena, Abriola)
61	499.9	148382	0.002	0.603	(Bionaz, Mondavio)
62	502.6	191420	0.003	0.606	(Etroubles, Cantiano)
63	504.7	144616	0.002	0.608	(Statte, Fano)
64	507.2	179144	0.003	0.611	(Peglio, Pisticci)
65	509.4	152160	0.002	0.613	(Cassiglio, Montefino)
66	511.4	139900	0.002	0.616	(Borgo Pace, Praia a Mare)
67	513.3	134538	0.002	0.618	(Bibbiena, Cirigliano)
68	515.8	173082	0.003	0.621	(Galeata, Marsicovetere)
69	517.7	139306	0.002	0.623	(Greve in Chianti, Castelmezzano)
70	520.2	173528	0.003	0.626	(Lizzano, Mondolfo)
71	522.4	149706	0.002	0.628	(Ischia di Castro, Montemesola)
72	525.0	182560	0.003	0.631	(Torricella, Mondolfo)
73	527.0	135842	0.002	0.633	(Gradara, San Paolo Albanese)
74	529.2	155050	0.002	0.636	(San Giovanni Valdarno, Miglionico)
75	532.0	193266	0.003	0.639	(Cellere, Lizzano)
76	534.8	193452	0.003	0.642	(Greve in Chianti, Cirigliano)
77	537.6	189660	0.003	0.645	(Monteiasi, Tavullia)
78	539.5	133562	0.002	0.647	(Morgex, Fabriano)
79	542.4	196648	0.003	0.650	(Cellere, Maruggio)
80	544.6	156582	0.003	0.653	(Torrita di Siena, Mongrassano)
81	547.1	165624	0.003	0.655	(Colli al Metauro, Fuscaldo)
82	549.2	146582	0.002	0.658	(Poggibonsi, Aliano)
83	551.2	135322	0.002	0.660	(Colli al Metauro, San Benedetto Ullano)
84	553.5	156334	0.003	0.662	(Piombino, Grumo Appula)
85	555.8	158972	0.003	0.665	(Sarteano, Torricella)
86	558.3	167494	0.003	0.668	(Tuoro sul Trasimeno, Zumpano)
87	560.3	132346	0.002	0.670	(Bettona, Pianopoli)
88	562.6	159984	0.003	0.672	(Cesena, Noepoli)
89	564.8	144134	0.002	0.675	(Terricciola, Corleto Perticara)
90	567.1	157634	0.003	0.677	(Forlimpopoli, Tursi)
91	569.3	144650	0.002	0.680	(Roccalbegna, Longobucco)
92	571.8	167254	0.003	0.682	(Bagno di Romagna, Statte)

Continued

Spatial Lag	h-distance (km)	Links	% Links	J(k)	Name of the Italian municipalities that determine the h-distance
93	573.7	127608	0.002	0.684	(Siena, Amendolara)
94	576.3	168020	0.003	0.687	(Crespina Lorenzana, Corleto Perticara)
95	578.4	137270	0.002	0.689	(Sava, Pennabilli)
96	581.3	187026	0.003	0.692	(Forlimpopoli, Nova Siri)
97	583.2	129670	0.002	0.694	(Bagno di Romagna, Taranto)
98	587.4	266656	0.004	0.699	(Nepi, Lentini)
99	591.7	282090	0.005	0.703	(Collevecchio, Lentini)
100	597.6	376858	0.006	0.709	(Ronciglione, Lentini)
101	602.3	298352	0.005	0.714	(Caprarola, Lentini)
102	607.4	320232	0.005	0.719	(Vasanello, Lentini)
103	612.7	339328	0.005	0.725	(Bassano in Teverina, Lentini)
104	618.1	337658	0.005	0.730	(Viterbo, Lentini)
105	626.8	548464	0.009	0.739	(Ladispoli, Pozzallo)
106	635.8	555474	0.009	0.748	(Scandriglia, Pachino)
107	644.8	558236	0.009	0.757	(Anguillara Sabazia, Pachino)
108	654.7	605934	0.010	0.766	(Monterosi, Pachino)
109	664.2	565212	0.009	0.776	(Civitavecchia, Pachino)
110	672.8	493060	0.008	0.783	(Villa San Giovanni in Tuscia, Pachino)
111	682.5	532642	0.009	0.792	(Soriano nel Cimino, Pachino)
112	691.6	473420	0.008	0.800	(San Gemini, Pachino)
113	701.5	487816	0.008	0.807	("Civitella dAgliano", Pachino)
114	712.1	495336	0.008	0.815	(Capalbio, Pachino)
115	721.5	412406	0.007	0.822	(Onano, Pachino)
116	729.5	337244	0.005	0.827	(Allerona, Pachino)
117	740.2	421080	0.007	0.834	(Scansano, Pachino)
118	749.5	347336	0.006	0.840	(Arcidosso, Pachino)
119	758.3	310714	0.005	0.845	(Castiglione del Lago, Pachino)
120	768.4	341190	0.005	0.850	(Castiglione della Pescaia, Pachino)
121	780.0	374746	0.006	0.856	(Portoferraio, Pachino)
122	790.1	313472	0.005	0.861	(Massa Marittima, Pachino)
123	800.9	326350	0.005	0.867	(Monterotondo Marittimo, Pachino)
124	810.8	296858	0.005	0.871	(Predoi, Lotzorai)
125	820.9	305822	0.005	0.876	(Predoi, Ortueri)
126	831.3	311214	0.005	0.881	(Cecina, Pachino)
127	842.0	318078	0.005	0.886	(Predoi, Villaurbana)
128	852.5	313336	0.005	0.891	(Predoi, Setzu)
129	863.3	329662	0.005	0.897	(Santa Maria a Monte, Pachino)
130	873.9	338908	0.005	0.902	(Predoi, Sanluri)
131	884.6	353692	0.006	0.908	(Predoi, Gonnosfanadiga)
132	894.5	340330	0.005	0.913	(San Benedetto Val di Sambro, Pachino)
133	905.2	386932	0.006	0.919	(Fabbriche di Vergemoli, Pachino)

Continued

Spatial Lag	h-distance (km)	Links	% Links	J(k)	Name of the Italian municipalities that determine the h-distance
134	915.5	392128	0.006	0.926	(Pieve Fosciana, Pachino)
135	926.4	425120	0.007	0.933	(Predoi, Santadi)
136	936.8	414170	0.007	0.939	(Predoi, Calasetta)
137	947.6	430800	0.007	0.946	(Pignone, Pachino)
138	957.2	376818	0.006	0.952	(Bomporto, Pachino)
139	968.1	408870	0.007	0.959	(Maissana, Pachino)
140	979.4	407604	0.007	0.965	(Leivi, Pachino)
141	990.4	370348	0.006	0.971	("Santo Stefano dAveto", Pachino)
142	1002.2	360116	0.006	0.977	(Pompeiana, Pachino)
143	1014.6	337812	0.005	0.983	(Bajardo, Pachino)
144	1026.2	270758	0.004	0.987	(Garessio, Pachino)
145	1037.9	225794	0.004	0.991	(Malvicino, Pachino)
146	1049.9	180950	0.003	0.993	(Niella Belbo, Pachino)
147	1061.7	133604	0.002	0.996	(Montelupo Albese, Pachino)
148	1073.6	94374	0.002	0.997	(Caraglio, Pachino)
149	1085.5	62782	0.001	0.998	(Lagnasco, Pachino)
150	1097.5	44050	0.001	0.999	(Rifreddo, Pachino)
151	1109.7	30902	0.000	0.999	(Bellino, Pachino)
152	1121.7	19482	0.000	1.000	(Frossasco, Pachino)
153	1133.3	11220	0.000	1.000	(San Francesco al Campo, Pachino)
154	1145.0	6466	0.000	1.000	(Germagnano, Pachino)
155	1157.2	3808	0.000	1.000	(Salbertrand, Pachino)
156	1168.2	1896	0.000	1.000	(Valprato Soana, Pachino)
157	1179.4	932	0.000	1.000	(Cogne, Pachino)
158	1191.0	492	0.000	1.000	(Gressan, Pachino)
159	1201.7	196	0.000	1.000	(Gignod, Pachino)
160	1213.6	80	0.000	1.000	(Prè Saint-Didier, Pachino)
161	1219.7	12	0.000	1.000	(Courmayeur, Pachino)
162	1224.0	2	0.000	1.000	(Courmayeur, Portopalo di Capo Passero)

6.1.2. CS Distance

Number of municipalities = 7,890 – Total links = 62,244,210

Spatial lag	h-distance (km)	Links	% Links	J(k)
1	16.3	269852	0.004	0.004
2	32.6	707810	0.011	0.016
3	48.9	1037740	0.017	0.032
4	65.2	1285378	0.021	0.053
5	81.5	1456148	0.023	0.076
6	97.8	1576398	0.025	0.102
7	114.1	1645254	0.026	0.128

Continued

Spatial lag	h-distance (km)	Links	% Links	J(k)
8	130.4	1674670	0.027	0.155
9	146.7	1670984	0.027	0.182
10	163.0	1643270	0.026	0.208
11	179.3	1585280	0.025	0.234
12	195.6	1509674	0.024	0.258
13	211.9	1439704	0.023	0.281
14	228.2	1363314	0.022	0.303
15	244.5	1298370	0.021	0.324
16	260.8	1244432	0.020	0.344
17	277.1	1200846	0.019	0.363
18	293.4	1168032	0.019	0.382
19	309.6	1144248	0.018	0.400
20	325.9	1115892	0.018	0.418
21	342.2	1089716	0.018	0.436
22	358.5	1059130	0.017	0.453
23	374.8	1029234	0.017	0.469
24	391.1	1015950	0.016	0.486
25	407.4	1026014	0.016	0.502
26	423.7	1051418	0.017	0.519
27	440.0	1071476	0.017	0.536
28	456.3	1093704	0.018	0.554
29	472.6	1123552	0.018	0.572
30	488.9	1146420	0.018	0.590
31	505.2	1147878	0.018	0.609
32	521.5	1143024	0.018	0.627
33	537.8	1126598	0.018	0.645
34	554.1	1115188	0.018	0.663
35	570.4	1097028	0.018	0.681
36	586.7	1066738	0.017	0.698
37	603.0	1043360	0.017	0.715
38	619.3	1029782	0.017	0.731
39	635.6	1017356	0.016	0.748
40	651.9	1004440	0.016	0.764
41	668.2	965410	0.016	0.779
42	684.5	900172	0.014	0.794
43	700.8	820590	0.013	0.807
44	717.1	753354	0.012	0.819
45	733.4	683142	0.011	0.830
46	749.7	620716	0.010	0.840
47	766.0	564658	0.009	0.849
48	782.3	524888	0.008	0.857
49	798.6	499496	0.008	0.865
50	814.9	491084	0.008	0.873
51	831.2	489564	0.008	0.881
52	847.5	483856	0.008	0.889
53	863.8	495646	0.008	0.897
54	880.1	526450	0.008	0.905

Continued

Spatial lag	h-distance (km)	Links	% Links	J(k)
55	896.4	557884	0.009	0.914
56	912.6	602946	0.010	0.924
57	928.9	637266	0.010	0.934
58	945.2	649362	0.010	0.945
59	961.5	634632	0.010	0.955
60	977.8	597054	0.010	0.964
61	994.1	541932	0.009	0.973
62	1010.4	472068	0.008	0.981
63	1026.7	390872	0.006	0.987
64	1043.0	297590	0.005	0.992
65	1059.3	208124	0.003	0.995
66	1075.6	130388	0.002	0.997
67	1091.9	75550	0.001	0.998
68	1108.2	46484	0.001	0.999
69	1124.5	25958	0.000	1.000
70	1140.8	12398	0.000	1.000
71	1157.1	5718	0.000	1.000
72	1173.4	2452	0.000	1.000
73	1189.7	878	0.000	1.000
74	1206.0	264	0.000	1.000
75	1222.3	60	0.000	1.000
76	1238.6	2	0.000	1.000

6.2. SGI Python Routine

To compute the Spatial Gini Index (SGI), an ad-hoc library was developed and implemented in Python to allow the immediate use of the new indicator. The ease of use and operational flexibility of this language are the main features of the routine. The library is particularly fast because it uses the well-known “NumPy” (Harris et al. 2020) and “SciPy” (Virtanen et al. 2020) libraries which, thanks to the vectorization of the functions, guarantee a high execution speed and represent the standard in matrix numerical calculation. One of the most famous and high performing libraries for Exploratory Spatial Data Analysis is certainly the PySAL library (Python Spatial Analysis Library) (Rey and Anselin 2010). Nevertheless, for efficiency reasons required by our use case it was necessary to create some “in-house” modules in order to better implement the SGI algorithm. The *k*-partitions obtained are also used for the spatial decomposition of the Gini index proposed by Rey and Smith (2013).

The principle of operation of the library is as follows:

The library can be downloaded as a Python package (Pirrotta 2022).

1. The SGI class accepts as input the geographic points (optionally the labels), the target variable and the type of partitioning (MM approach is considered here) – the geographic points are the centroids of the territorial units considered,
2. The Euclidean distance matrix is calculated,
3. The variability matrix is calculated,

4. The algorithm finds all the *n* * (*n*-1) connections between the *n* territorial units. This process takes place inside a loop,
5. According to the MM method, for each territorial unit the minimum distance with all the other units is taken into consideration. Among these *n* minimum values, the maximum is taken. This value represents the *h*-distance,
6. For each iteration, only connections inside the range distance between minimum threshold (previous h-distance) and maximum threshold (current h-distance) are selected,
7. For each iteration the variability matrix is multiplied element-wise by the spatial weight matrix generated in point 6. By adding the values obtained, the total variability for each spatial-lag is obtained,
8. In order to calculate the successive *h*-distances according to the SGI approach, at each iteration it is necessary to filter from the Euclidian distance matrix the distances between units less than or equal to the previous *h*-distance. and
9. When all the units are connected, the process ends with the calculation of the index. The result obtained is the Spatial Gini Index (SGI).

The flow chart of the procedure is presented in the figure below.

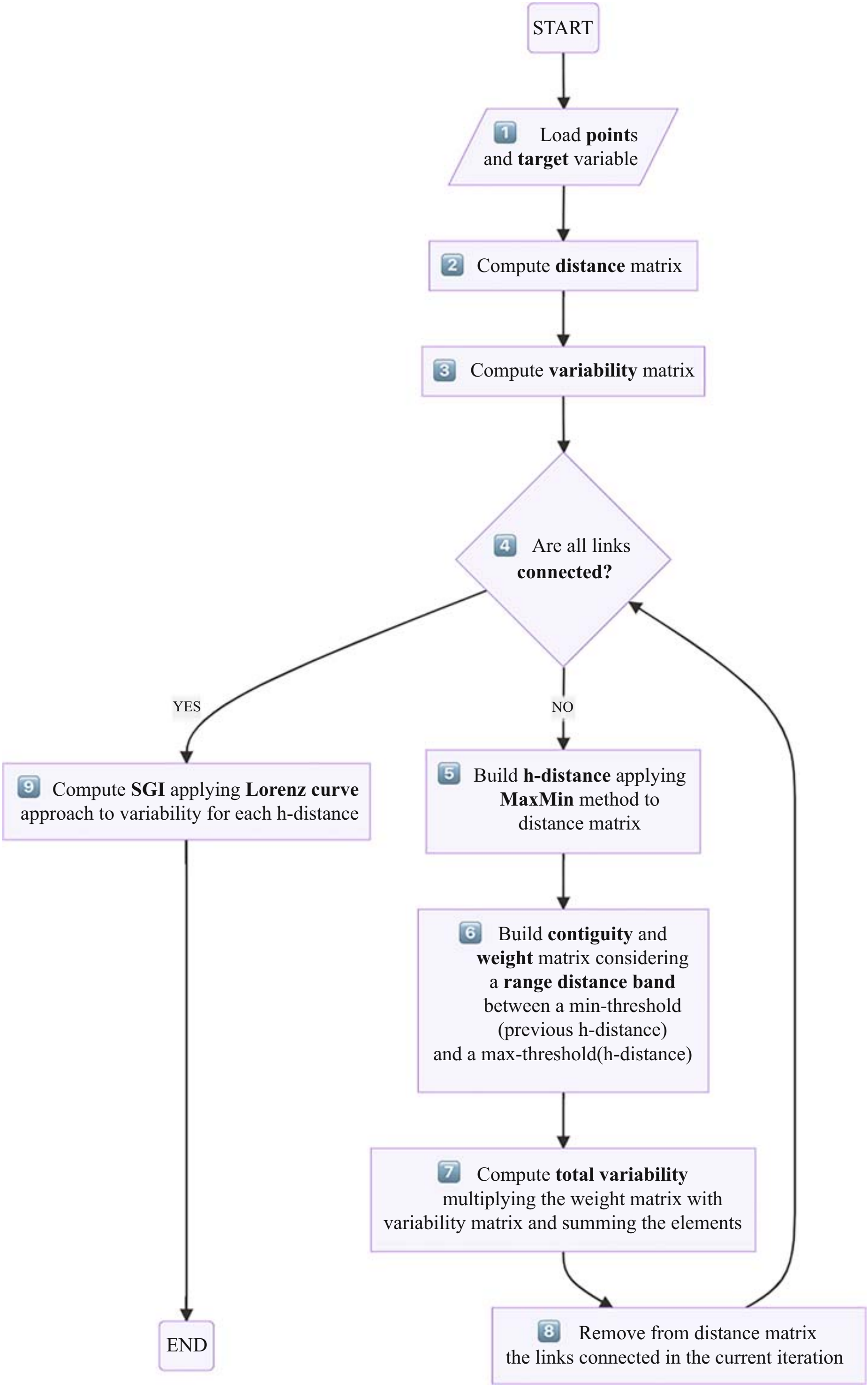
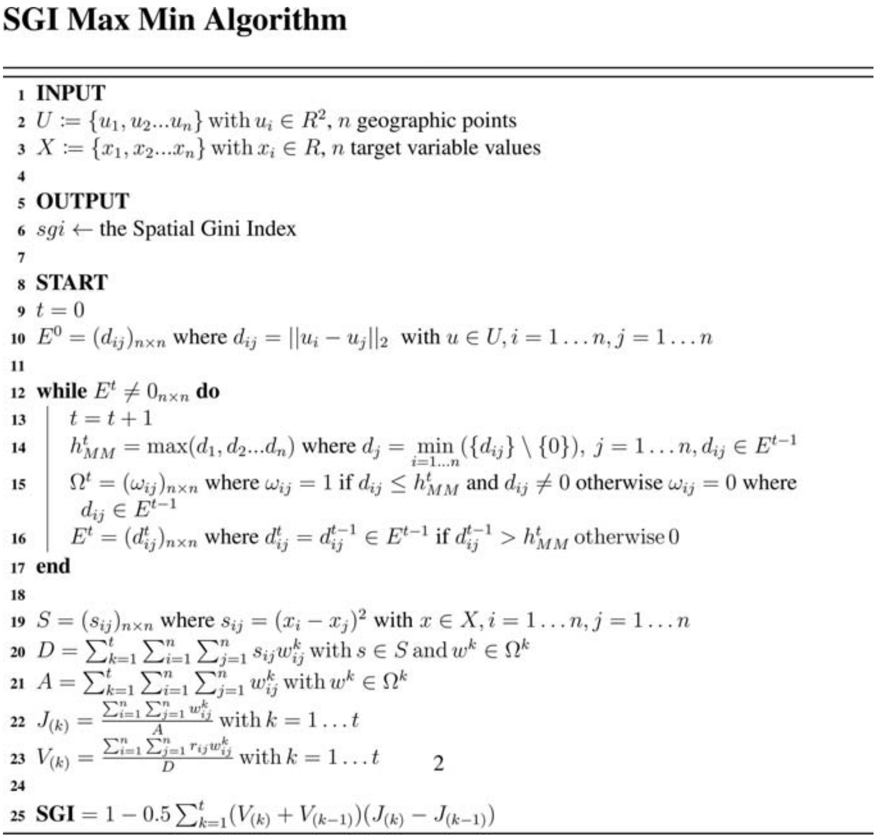

START
1 Load points and target variable
2 Compute distance matrix
3 Compute variability matrix
4 Are all links connected?
YES
NO
9 Compute SGI applying Lorenz curve approach to variability for each h-distance
5 Build h-distance applying MaxMin method to distance matrix
6 Build contiguity and weight matrix considering a range distance band between a min-threshold (previous h-distance) and a max-threshold(h-distance)
7 Compute total variability multiplying the weight matrix with variability matrix and summing the elements
8 Remove from distance matrix the links connected in the current iteration
END

Fig. 13. Flow chart of the procedure for the computation of SGI.

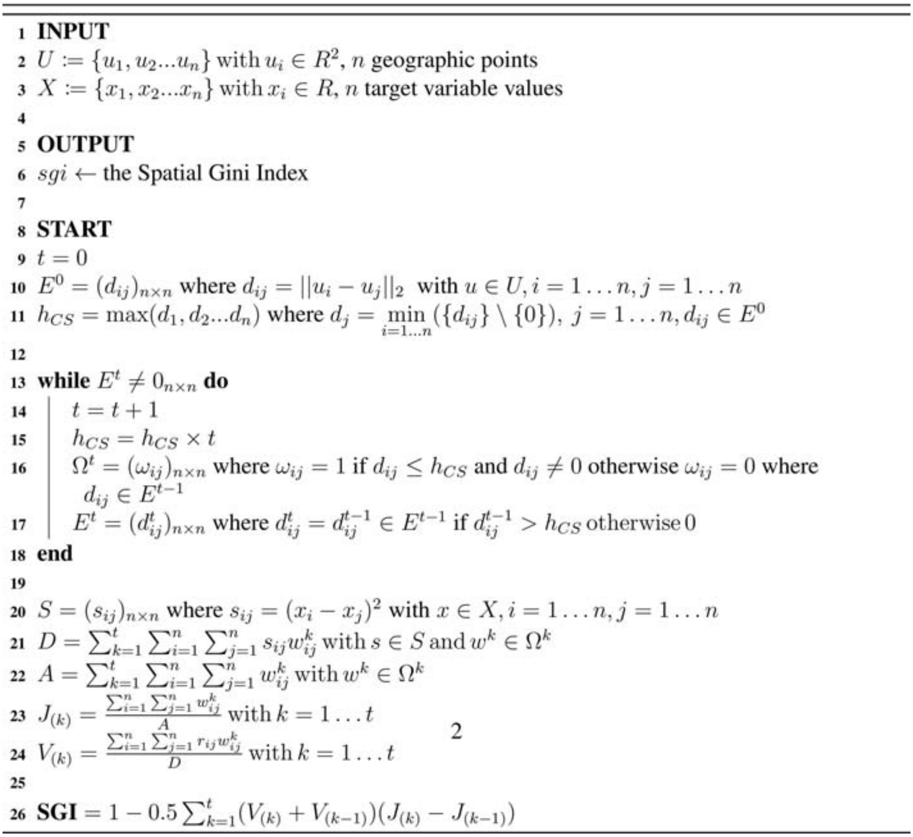
6.3. Algorithm and Pseudo Code



SGI Max Min Algorithm

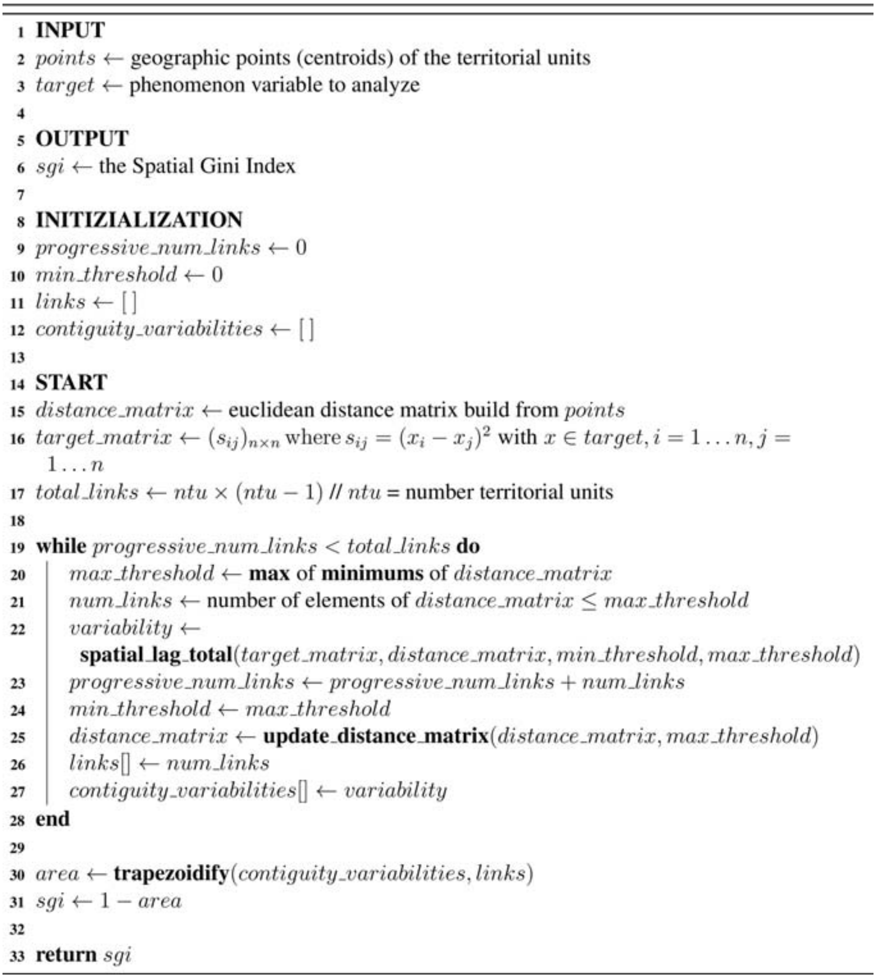
```
INPUT
$U := \{u_1, u_2 ... u_n\}$ with $u_i \in R^2$, $n$ geographic points
$X := \{x_1, x_2 ... x_n\}$ with $x_i \in R$, $n$ target variable values

OUTPUT
$sgi \leftarrow$ the Spatial Gini Index

START
$t = 0$
$E^0 = (d_{ij})_{n \times n}$ where $d_{ij} = ||u_i - u_j||_2$ with $u \in U, i = 1 \dots n, j = 1 \dots n$

while $E^t \neq 0_{n \times n}$ do
    $t = t + 1$
    $h^t_{MM} = \max(d_1, d_2 ... d_n)$ where $d_j = \min_{i=1...n}(\{d_{ij}\} \setminus \{0\}),\ j = 1 \dots n, d_{ij} \in E^{t-1}$
    $\Omega^t = (\omega_{ij})_{n \times n}$ where $\omega_{ij} = 1$ if $d_{ij} \leq h^t_{MM}$ and $d_{ij} \neq 0$ otherwise $\omega_{ij} = 0$ where
      $d_{ij} \in E^{t-1}$
    $E^t = (d^t_{ij})_{n \times n}$ where $d^t_{ij} = d^{t-1}_{ij} \in E^{t-1}$ if $d^{t-1}_{ij} > h^t_{MM}$ otherwise 0
end

$S = (s_{ij})_{n \times n}$ where $s_{ij} = (x_i - x_j)^2$ with $x \in X, i = 1 \dots n, j = 1 \dots n$
$D = \sum_{k=1}^{t} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} w^k_{ij}$ with $s \in S$ and $w^k \in \Omega^k$
$A = \sum_{k=1}^{t} \sum_{i=1}^{n} \sum_{j=1}^{n} w^k_{ij}$ with $w^k \in \Omega^k$
$J_{(k)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w^k_{ij}}{A}$ with $k = 1 \dots t$
$V_{(k)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij} w^k_{ij}}{D}$ with $k = 1 \dots t$        2

SGI $= 1 - 0.5 \sum_{k=1}^{t} (V_{(k)} + V_{(k-1)})(J_{(k)} - J_{(k-1)})$
```

SGI Constant Step Algorithm



```
INPUT
$U := \{u_1, u_2 ... u_n\}$ with $u_i \in R^2$, $n$ geographic points
$X := \{x_1, x_2 ... x_n\}$ with $x_i \in R$, $n$ target variable values

OUTPUT
$sgi \leftarrow$ the Spatial Gini Index

START
$t = 0$
$E^0 = (d_{ij})_{n \times n}$ where $d_{ij} = ||u_i - u_j||_2$ with $u \in U, i = 1 \dots n, j = 1 \dots n$
$h_{CS} = \max(d_1, d_2 ... d_n)$ where $d_j = \min_{i=1 \dots n} (\{d_{ij}\} \setminus \{0\}),\ j = 1 \dots n, d_{ij} \in E^0$

while $E^t \neq 0_{n \times n}$ do
    $t = t + 1$
    $h_{CS} = h_{CS} \times t$
    $\Omega^t = (\omega_{ij})_{n \times n}$ where $\omega_{ij} = 1$ if $d_{ij} \leq h_{CS}$ and $d_{ij} \neq 0$ otherwise $\omega_{ij} = 0$ where
      $d_{ij} \in E^{t-1}$
    $E^t = (d^t_{ij})_{n \times n}$ where $d^t_{ij} = d^{t-1}_{ij} \in E^{t-1}$ if $d^{t-1}_{ij} > h_{CS}$ otherwise 0
end

$S = (s_{ij})_{n \times n}$ where $s_{ij} = (x_i - x_j)^2$ with $x \in X, i = 1 \dots n, j = 1 \dots n$
$D = \sum_{k=1}^{t} \sum_{i=1}^{n} \sum_{j=1}^{n} s_{ij} w^k_{ij}$ with $s \in S$ and $w^k \in \Omega^k$
$A = \sum_{k=1}^{t} \sum_{i=1}^{n} \sum_{j=1}^{n} w^k_{ij}$ with $w^k \in \Omega^k$
$J_{(k)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w^k_{ij}}{A}$ with $k = 1 \dots t$
                                        2
$V_{(k)} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} r_{ij} w^k_{ij}}{D}$ with $k = 1 \dots t$

SGI $= 1 - 0.5 \sum_{k=1}^{t} (V_{(k)} + V_{(k-1)})(J_{(k)} - J_{(k-1)})$
```

SGI MaxMin Pseudocode



```
1  INPUT
2  points ← geographic points (centroids) of the territorial units
3  target ← phenomenon variable to analyze
4
5  OUTPUT
6  sgi ← the Spatial Gini Index
7
8  INITIZIALIZATION
9  progressive_num_links ← 0
10 min_threshold ← 0
11 links ← [ ]
12 contiguity_variabilities ← [ ]
13
14 START
15 distance_matrix ← euclidean distance matrix build from points
16 target_matrix ← $(s_{ij})_{n \times n}$ where $s_{ij} = (x_i - x_j)^2$ with $x \in target, i = 1 \dots n, j = 1 \dots n$
17 total_links ← ntu × (ntu − 1) // ntu = number territorial units
18
19 while progressive_num_links < total_links do
20     max_threshold ← max of minimums of distance_matrix
21     num_links ← number of elements of distance_matrix ≤ max_threshold
22     variability ←
         spatial_lag_total(target_matrix, distance_matrix, min_threshold, max_threshold)
23     progressive_num_links ← progressive_num_links + num_links
24     min_threshold ← max_threshold
25     distance_matrix ← update_distance_matrix(distance_matrix, max_threshold)
26     links[] ← num_links
27     contiguity_variabilities[] ← variability
28 end
29
30 area ← trapezoidify(contiguity_variabilities, links)
31 sgi ← 1 − area
32
33 return sgi
```

6.2.1. Pseudocode Explanation

The library was developed using programming techniques and strategies to ensure speed, robustness, scalability, high performance, and efficiency. Below the explanation of the algorithm.

2−3: The algorithm accepts as input the geographic **points** (centroids of the territorial units considered) and the **target** variable

15: the Euclidean **distance matrix** is calculated

16: the **target matrix** is calculated

17: the algorithm finds all the $n \times (n - 1)$ connections between the n territorial units

19: the loop ends when **all links** between territorial units are **connected**

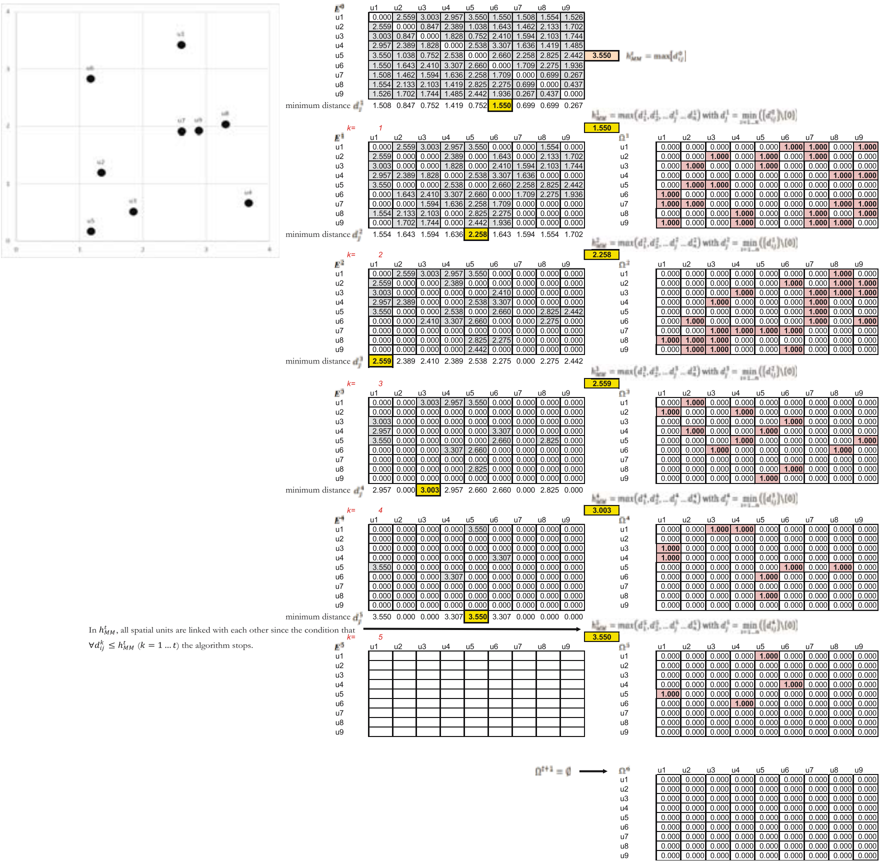
20: Starting from the **distance matrix** (dim. $ntu \times ntu$), for each territorial unit the **minimum** distance with all the other units is taken into consideration. Among these n

minimum values (array), the **maximum** (scalar) is taken. This value represents both the max **threshold** and the current **h distance**
21: for each iteration the number of links corresponds to the number of values in the distance matrix less or equal to the maximum threshold (current h distance);
22: In the **spatial_lag_total** procedure a spatial boolean weights matrix is created (RangeDistance-Band object) selecting connections between minimum and maximum threshold. The target matrix is so multiplied **element-wise** by the spatial weight matrix. By adding the values obtained, the **total variability** for each spatial-lag is obtained;
25: for each iteration we update the distance matrix removing (reset to zero) values less or equal to the maximum threshold;
30: for the estimation of **Gini coefficients** we calculate the **Lorenz curve** area using the **trapezium rule**
31: the **Spatial Gini Index** is obtained subtracting area from 1

6.3. Summary of the Results

Years	Italian population				Foreign population			
	MM		CS		MM		CS	
	G	SGI	G	SGI	G	SGI	G	SGI
2002	0.712	0.485	0.712	0.485	0.802	0.446	0.802	0.444
2010	0.710	0.485	0.710	0.485	0.782	0.448	0.782	0.447
2018	0.715	0.483	0.715	0.483	0.797	0.450	0.797	0.449

6.4. Toy Example MaxMin Distance Method Procedure (Simulated Nine Spatial Units)



E^0	u1	u2	u3	u4	u5	u6	u7	u8	u9
u1	0.000	2.559	3.003	2.957	3.550	1.550	1.508	1.554	1.526
u2	2.559	0.000	0.847	2.389	1.038	1.643	1.462	2.133	1.702
u3	3.003	0.847	0.000	1.828	0.752	2.410	1.594	2.103	1.744
u4	2.957	2.389	1.828	0.000	2.538	3.307	1.636	1.419	1.485
u5	3.550	1.038	0.752	2.538	0.000	2.660	2.258	2.825	2.442
u6	1.550	1.643	2.410	3.307	2.660	0.000	1.709	2.275	1.936
u7	1.508	1.462	1.594	1.636	2.258	1.709	0.000	0.699	0.267
u8	1.554	2.133	2.103	1.419	2.825	2.275	0.699	0.000	0.437
u9	1.526	1.702	1.744	1.485	2.442	1.936	0.267	0.437	0.000
minimum distance d_j^1	1.508	0.847	0.752	1.419	0.752	1.550	0.699	0.699	0.267

3.550 $h^0_{MM} = \max[d^0_{ij}]$

$h^1_{MM} = \max(d^1_1, d^1_2, \dots d^1_j \dots d^1_n)$ with $d^1_j = \min_{i=1\dots n}(\{d^0_{ij}\}\setminus\{0\})$

k= *1* 1.550

E^1	u1	u2	u3	u4	u5	u6	u7	u8	u9
u1	0.000	2.559	3.003	2.957	3.550	0.000	0.000	1.554	0.000
u2	2.559	0.000	0.000	2.389	0.000	1.643	0.000	2.133	1.702
u3	3.003	0.000	0.000	1.828	0.000	2.410	1.594	2.103	1.744
u4	2.957	2.389	1.828	0.000	2.538	3.307	1.636	0.000	0.000
u5	3.550	0.000	0.000	2.538	0.000	2.660	2.258	2.825	2.442
u6	0.000	1.643	2.410	3.307	2.660	0.000	1.709	2.275	1.936
u7	0.000	0.000	1.594	1.636	2.258	1.709	0.000	0.000	0.000
u8	1.554	2.133	2.103	0.000	2.825	2.275	0.000	0.000	0.000
u9	0.000	1.702	1.744	0.000	2.442	1.936	0.000	0.000	0.000
minimum distance d_j^2	1.554	1.643	1.594	1.636	2.258	1.643	1.594	1.554	1.702

Ω^1	u1	u2	u3	u4	u5	u6	u7	u8	u9
u1	0.000	0.000	0.000	0.000	0.000	1.000	1.000	0.000	1.000
u2	0.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000	0.000
u3	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
u4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000
u5	0.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
u6	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u7	1.000	1.000	0.000	0.000	0.000	0.000	0.000	1.000	1.000
u8	0.000	0.000	0.000	1.000	0.000	0.000	1.000	0.000	1.000
u9	1.000	0.000	0.000	1.000	0.000	0.000	1.000	1.000	0.000

$h^2_{MM} = \max(d^2_1, d^2_2, \dots d^2_j \dots d^2_n)$ with $d^2_j = \min_{i=1\dots n}(\{d^1_{ij}\}\setminus\{0\})$

k= *2* 2.258

E^2	u1	u2	u3	u4	u5	u6	u7	u8	u9
u1	0.000	2.559	3.003	2.957	3.550	0.000	0.000	0.000	0.000
u2	2.559	0.000	0.000	2.389	0.000	0.000	0.000	0.000	0.000
u3	3.003	0.000	0.000	0.000	0.000	2.410	0.000	0.000	0.000
u4	2.957	2.389	0.000	0.000	2.538	3.307	0.000	0.000	0.000
u5	3.550	0.000	0.000	2.538	0.000	2.660	0.000	2.825	2.442
u6	0.000	0.000	2.410	3.307	2.660	0.000	0.000	2.275	0.000
u7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u8	0.000	0.000	0.000	0.000	2.825	2.275	0.000	0.000	0.000
u9	0.000	0.000	0.000	0.000	2.442	0.000	0.000	0.000	0.000
minimum distance d_j^3	2.559	2.389	2.410	2.389	2.538	2.275	0.000	2.275	2.442

Ω^2	u1	u2	u3	u4	u5	u6	u7	u8	u9
u1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000
u2	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	1.000
u3	0.000	0.000	0.000	1.000	0.000	0.000	1.000	1.000	1.000
u4	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	0.000
u5	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000
u6	0.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000
u7	0.000	0.000	1.000	1.000	1.000	1.000	0.000	0.000	0.000
u8	1.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000
u9	0.000	1.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000

$h^3_{MM} = \max(d^3_1, d^3_2, \dots d^3_j \dots d^3_n)$ with $d^3_j = \min_{i=1\dots n}(\{d^2_{ij}\}\setminus\{0\})$

k= *3* 2.559

E^3	u1	u2	u3	u4	u5	u6	u7	u8	u9
u1	0.000	0.000	3.003	2.957	3.550	0.000	0.000	0.000	0.000
u2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u3	3.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u4	2.957	0.000	0.000	0.000	0.000	3.307	0.000	0.000	0.000
u5	3.550	0.000	0.000	0.000	0.000	2.660	0.000	2.825	0.000
u6	0.000	0.000	0.000	3.307	2.660	0.000	0.000	0.000	0.000
u7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u8	0.000	0.000	0.000	0.000	2.825	0.000	0.000	0.000	0.000
u9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
minimum distance d_j^4	2.957	0.000	3.003	2.957	2.660	2.660	0.000	2.825	0.000

Ω^3	u1	u2	u3	u4	u5	u6	u7	u8	u9
u1	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u2	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
u3	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
u4	0.000	1.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
u5	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	1.000
u6	0.000	0.000	1.000	0.000	0.000	0.000	0.000	1.000	0.000
u7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u8	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
u9	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000

$h^4_{MM} = \max(d^4_1, d^4_2, \dots d^4_j \dots d^4_n)$ with $d^4_j = \min_{i=1\dots n}(\{d^3_{ij}\}\setminus\{0\})$

k= *4* 3.003

E^4	u1	u2	u3	u4	u5	u6	u7	u8	u9
u1	0.000	0.000	0.000	0.000	3.550	0.000	0.000	0.000	0.000
u2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u4	0.000	0.000	0.000	0.000	0.000	3.307	0.000	0.000	0.000
u5	3.550	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u6	0.000	0.000	0.000	3.307	0.000	0.000	0.000	0.000	0.000
u7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
minimum distance d_j^5	3.550	0.000	0.000	3.307	3.550	3.307	0.000	0.000	0.000

Ω^4	u1	u2	u3	u4	u5	u6	u7	u8	u9
u1	0.000	0.000	1.000	1.000	0.000	0.000	0.000	0.000	0.000
u2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u3	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u4	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u5	0.000	0.000	0.000	0.000	0.000	1.000	0.000	1.000	0.000
u6	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
u7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u8	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
u9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

In h^t_{MM}, all spatial units are linked with each other since the condition that $\forall d^k_{ij} \le h^t_{MM}$ $(k = 1 \dots t)$ the algorithm stops.

$h^5_{MM} = \max(d^5_1, d^5_2, \dots d^5_j \dots d^5_n)$ with $d^5_j = \min_{i=1\dots n}(\{d^4_{ij}\}\setminus\{0\})$

k= *5* 3.550

E^5	u1	u2	u3	u4	u5	u6	u7	u8	u9
u1									
u2									
u3									
u4									
u5									
u6									
u7									
u8									
u9									

Ω^5	u1	u2	u3	u4	u5	u6	u7	u8	u9
u1	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000
u2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u4	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
u5	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u6	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000
u7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

$\Omega^{t+1} = \emptyset$ →

Ω^6	u1	u2	u3	u4	u5	u6	u7	u8	u9
u1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u2	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u3	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u4	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u5	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u6	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u8	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
u9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

7. References

Alleva, G. 1987. “Autocorrelazione spaziale nel caso di contiguità di ordine superiore al primo.” In Proceedings of the Italian Statistical Society Conference “Informazione ed analisi statistica per aree regionali e sub regionali”, Galeno Editrice, Perugia, Italia: 1–10.

Anselin, L. 1999. “The Future of Spatial Analysis in the Social Sciences.” *Geographic Information Sciences* 5(2): 67–76. DOI: https://doi.org/10.1080/10824009909480516.

Arbia, G. 2001. “The role of spatial effects in the empirical analysis of regional concentration.” *Journal of Geographical Systems* 3(3): 271–281. DOI: https://doi.org/10.1007/PL00011480.

Arbia, G., R. Benedetti, and G. Espa. 1996. “Effect of MAUP on image classification.” *Geographical System* 3: 123–141.

Arbia, G., and G. Piras. 2009. “A new class of spatial concentration measures.” *Computational Statistics and Data Analysis* 53(21): 4471–4481. DOI: https://doi.org/10.1016/j.csda.2009.07.003.

Benassi, F., F. Lipizzi, and S. Strozza. 2019. “Detecting foreigners’ spatial residential patterns in urban contexts: two tales from Italy.” *Applied Spatial Analysis and Policy* 12(2): 301–319. DOI: https://doi.org/10.1007/s12061-017-9243-5.

Benassi, F., R. Iglesias-Pascual, and L. Salvati. 2020. “Residential segregation and social diversification: Exploring spatial settlement patterns of foreign population in Southern European cities.” *Habitat International* 101:102200. DOI: https://doi.org/10.1016/j.habitatint.2020.102200.

Bonferroni, C.E. 1938. *Elementi di Statistica Generale*. Litografai Feilice Gili: Torino.

Brown, L.A., and S.Y. Chung. 2006. “Spatial segregation, segregation indices and the geographical perspectives.” *Population, Space and Place* 12(2): 125–143. DOI: https://doi.org/10.1002/psp.403.

Crespo, R., and I. Hernandez. 2020. “On the spatially explicit Gini coefficient: The case study of Chile – a high-income developing country.” *Letters in Spatial and Resource Sciences* 13(1): 37–47. DOI: https://doi.org/10.1007/s12076-020-00243-4.

Dawkins, C.J. 2004. “Measuring the spatial pattern of residential segregation.” *Urban Studies* 41(4): 833–851. DOI: https://doi.org/10.1080/0042098042000194.

Dawkins, C.J. 2006. “The spatial pattern of black-white segregation in US metropolitan areas: An exploratory analysis.” *Urban Studies* 43(11): 1943–1969. DOI: https://doi.org/10.1080/00420980600897792.

Eurostat. 2015. *Geospatial analysis at Eurostat*. Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Geospatial_analysis_at_Eurostat#Spatial_analysis_for_statistical_purposes.

Fotheringham A.S. 1997. “Trends in Quantitative Methods I: Stressing the Local.” *Progress in Human Geography* 21(1): 88–96. DOI: https://doi.org/10.1191/030913297676693207.

Frosini, B.V. 1996. “Recenti contributi Italiani alla statistica descrittiva.” *Statistica Applicata* 8(2): 361–396.

Gini, C. 1912. “Variabilità e mutabilità. Contributo allo studio delle distribuzioni e delle relazioni statistiche.” *Studi economico-giuridici, Anno III, parte II. Facoltà di giurisprudenza della Regia Università di Cagliari*, Cuppini, Bologna.

Gini, C. 1914. “Sulla misura della concentrazione della variabilità dei caratteri.” *Atti del reale istituto Veneto di scienze, lettere e arti* 73(2): 1203–1248.

Gini, C. 1921. “Measurement of inequality of incomes.” *Economic Journal* 31: 124–126. DOI: https://doi.org/10.2307/2223319.

Gini, C. 2005. “On the measurement of concentration and variability of characters.” *Metron* 63(1): 3–38. Available at: https://www.dss.uniroma1.it/RePec/mtn/articoli/2005-1-1.pdf.

Harris, C.R., K.J. Millman, S.J. van der Walt. R. Gommers, P. Virtanen, David Cournapeau, E. Wieser, J. Taylor, S. Berg, N.J. Smith, R. Kern, M. Picus, S. Hoyer, M.H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T.E. Oliphan. 2020. “Array programming with NumPy.” *Nature* 585: 357–362. DOI: https://doi.org/10.1038/s41586-020-2649-2.

Hoover, E. 1941.“Interstate redistribution of population, 1850–1940.” *Journal of Economic History* 1: 199–205. DOI: https://doi.org/10.1017/S0022050700052980.

Howell, F.M., J.R. Porter, and S.A. Matthews. 2016.*Recapturing space: New middle-range theory in spatial demography*. Berlin: Springer. DOI: https://doi.org/10.1007/978-3-319-22810-5.

Livi Bacci, M. 1999. *The Population of Europe: A History*. Oxford: Blackwell

Leti G. 1983 *Statistica descrittiva*, Bologna: Il Mulino.

Logan J.R. 2012. “Making a Place for Space: Spatial Thinking in Social Science.” *Annual Review of Sociology* 38: 507–524. DOI: https://doi.org/10.1146/annurev-soc-071811-145531.

Lorenz, M.O. 1905. “Methods of measuring the concentration of wealth.” *Publications of the American Statistical Association* 9(70): 209–219. DOI: https://doi.org/10.2307/2276207.

Matthews, S.A., and D.M. Parker. 2013. “Progress in spatial demography.” *Demographic Research* 28: 271–312. DOI: https://doi.org/10.4054/DemRes.2013.28.10.

Massey, D.S., and N.A. Denton. 1988. “The dimensions of residential segregation.” *Social forces* 67(2): 281–315. DOI: https://doi.org/10.1093/sf/67.2.281.

Miller, H.J. 2004. “Tobler’s first law and spatial analysis.” *Annals of the association of American geographers* 94(2): 284–289. DOI: https://doi.org/10.1111/j.1467-8306.2004.09402005.x.

Mucciardi, M. (2008a). “Use of a flexible weight matrix in a local spatial statistic.” In: First joint meeting of the Société Francophone de Classification and the Classification and Data Analysis Group of the Italian Society of Statistics. 5: 385–388. Naples: Edizioni Scientifiche Italiane. DOI: https://doi.org/10.1007/978-3-642-13312-1.

Mucciardi, M. (2008b). “Geographic information and global index of spatial variability.” Proceedings of the VI International Conference on Stochastics Geometry, Convex Bodies, Empirical Measures. In *Rendiconti del Circolo dei Matematici di Palermo 80*: 241–250.

Mucciardi, M., and F. Benassi. 2023. “Measuring the spatial concentration of population. A new approach based on the graphical representation of the Gini index.” *Quality and Quantity*. DOI: https://doi.org/10.1007/s11135-022-01607-2.

Mucciardi, M., and P. Bertuccelli. 2007. “S-Joint: a new software for the analysis of spatial data”. In Proceedings of the Italian Statistical Society Annual Conference, Venice, 6–8 June: 647–648. Available at: http://sis2007.unive.it/.

Novelli, E. and S. Ocelli. 1999. “Describing spatial patterns’ profile: diversification measure.” *European Journal of Geography*. DOI: https://doi.org/10.4000/cybergeo.4930.

Panzera, D., and P. Postiglione. 2020. “Measuring the spatial dimension of regional inequality: An approach based on the Gini correlation measure.” *Social Indicators Research* 148(2): 379–394. DOI: https://doi.org/10.1007/s11205-019-02208-7.

Pearl, R., and L.J. Reed. 1920. “On the rate of growth of the population of the USA since 1970 and its mathematical representation.” *Proceedings of the National Academy of Science* 6: 275-288. DOI: https://doi.org/10.1073/pnas.6.6.275.

Pirrotta, G. 2022. *The Spatial Gini Index, a Python library to compute a new indicator to measure the spatial concentration*. Available at: https://github.com/gpirrotta/spatial-gini-index.

Reardon, S.F., and D. O'Sullivan. 2004. "Measures of spatial segregation." *Sociological Methodology* 34(1): 121–162. DOI: https://doi.org/10.1111/j.0081-1750.2004.00150.x.

Rey, S.J., and L. Anselin. 2010. "PySAL: A Python Library of Spatial Analytical Methods." In *Handbook of Applied Spatial Analysis,* edited by *M*. Fischer, and A. Getis. Springer, Berlin, Heidelberg. DOI: https://doi.org/10.1007/978-3-642-03647-7_11.

Rey, S.J., and R.J. Smith. 2013. "A spatial decomposition of the Gini coefficient." *Letters in Spatial and Resource Sciences* 6(2): 55–70. DOI: https://doi.org/10.1007/s12076-012-0086-z.

Strozza, S., F. Benassi, R. Ferrara, and G. Gallo. 2016. "Recent demographic trends in the major Italian urban agglomerations: The role of foreigners." *Spatial Demography* 4(1): 39–70. DOI: https://doi.org/10.1007/s40980-015-0012-2.

Tobler, W.R. 1970. "A computer movie simulating urban growth in the Detroit region." *Economic Geography* 46: 234–240. DOI: https://doi.org/10.2307/143141.

Verhulst, P.F. 1838. "Notice sur la loi que la population poursuit dans accroissement." *Correspondance Mathématique et Physique* 10: 113-121.

Voss, P.R. 2007. "Demography as a spatial social science." *Population research and policy review* 26(5): 457–476. DOI: https://doi.org/10.1007/s11113-007-9047-4

Wright, J.K. 1937. "Some measures of distributions." *Annals of the Association of American Geographers* 27(4): 177–211.

Virtanen, P., R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, İ. Polat, Y. Feng, E.W. Moore, and J. VanderPlas. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods*, 17: 261–272. DOI: https://doi.org/10.1038/s41592-019-0686-2.

Zenga, M. 1984. "Proposta per un indice di concentrazione basato sui rapporti fra quantili di popolazione e quantili di reddito." *Giornale degli Economisti e Annali di Economia* 43(5/6): 301–326. Available at: http://www.jstor.org/stable/23246074.

Zenga, M. 1985. "Un secondo indice di concentrazione basato sui rapporti fra quantili di popolazione e quantili di reddito." *Rivista di Statistica Applicata* 18: 143–154.

Received June 2022
Revised October 2022
Accepted January 2023

Journal of Official Statistics, Vol. 39, No. 3, 2023, pp. 325–349, http://dx.doi.org/10.2478/JOS-2023-0015

Predicting Days to Respondent Contact in Cross-Sectional Surveys Using a Bayesian Approach

Stephanie Coffey[1] *and Michael R. Elliott*[2]

Surveys estimate and monitor a variety of data collection parameters, including response propensity, number of contacts, and data collection costs. These parameters can be used as inputs to a responsive/adaptive design or to monitor the progression of a data collection period against predefined expectations. Recently, Bayesian methods have emerged as a method for combining historical information or external data with data from the in-progress data collection period to improve prediction. We develop a Bayesian method for predicting a measure of case-level progress or productivity, the estimated time lag, in days, between first contact attempt and first respondent contact. We compare the quality of predictions from the Bayesian method to predictions generated from more commonly-used predictive methods that leverage data from only historical data collection periods or the in-progress round of data collection. Using prediction error and misclassification as short- or long- day lags, we demonstrate that the Bayesian method results in improved predictions close to the day of the first contact attempt, when these predictions may be most informative for interventions or interviewer feedback. This application adds to evidence that combining historical and current information about data collection, in a Bayesian framework, can improve predictions of data collection parameters.

Key words: Adaptive design; hurdle models; national health interview survey.

1. Introduction

Making contact with a survey respondent is an important part of the data collection process. Without a successful contact, it is not only impossible to obtain cooperation and a completed interview, possibly introducing nonresponse bias, but it may also be impossible to determine the eligibility of a case, affecting response rates, nonresponse adjustments, and ultimately, variances of estimates. Noncontact makes up a substantial portion of nonresponse in face-to-face surveys. Durrant and Steele (2009) discuss six large,

[1] U.S. Census Bureau, Research and Methodology, 4600 Silver Hill Road, Suitland, Maryland, 20746 U.S.A. Email: stephmcoffey@gmail.com
[2] University of Michigan, Department of Biostatistics, School of Public Health, University of Michigan, M4041 SPH II, 1420 Washington Heights, Ann Arbor, Michigan 48109, Michigan, 48105 U.S.A. Email: mrelliot@umich.edu

Acknowledgments: Research reported in this publication was supported by the National Center for Science and Engineering Statistics of the National Science Foundation under contract number 0099-FY16-IPA-0002, and the National Institute on Aging of the National Institutes of Health under Award Number R01AG058599. This study design and analysis was not preregistered. The opinions and conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau, the National Science Foundation, or the National Institutes of Health. All results have been reviewed to ensure that no confidential information is disclosed. The authors thank the journal reviewers as well as Stephen Blumberg, James Dahlhamer, James Wagner, Brady West, and Yan Li for their review and helpful comments. The data used to prepare this manuscript are restricted to uses approved by the National Center for Health Statistics (NCHS). As a result, these data cannot be included as part of supplementary materials due to privacy and confidentiality requirements. However, researchers can apply to access this data in the NCHS Research Data Center (RDC) by following the proposal process outlined at this location: https://www.cdc.gov/rdc/index.htm.

government-sponsored, face-to-face household surveys carried out in the UK which have noncontact rates ranging from 8% to 40% of sample cases. Williams and Brick (2018) report that for nine government-sponsored, primarily face-to-face household surveys in the United States, noncontact rates as a portion of nonresponse range from four percent to 40% for the most recent year included in the article, 2014. Through a trend analysis across national statistical institutes, Luiten et al. (2020) also found evidence that noncontact rates are increasing in the United States, though evidence was mixed in other countries.

Groves and Couper (1998) discuss methods for reducing nonresponse due to noncontact, including planning for numerous contact attempts in face-to-face surveys, and requiring interviewers to make attempts early enough in the data collection period to make contact with all sample units. Waiting too long to begin making attempts could lead to nonresponse due to noncontact rather than post-contact reluctance or refusal. Biffignandi and Bethlehem (2021) note that an effective call management system can mitigate this problem, particularly in a centralized telephone environment, however, that solution may not be applicable to decentralized interviewing environments, such as in-person interviewer-administered surveys.

In order to improve survey outcomes, survey managers might want to monitor the expected time lag, in days, between a first attempt and first contact with a sample member to better understand the expected progress of data collection operations. If, for example, these day lags are longer than expected, or are leading to nonresponse due to noncontact, survey managers could attempt to reduce this lag by intervening with respect to individual cases shortly after a first contact attempt is made. Those interventions could reflect different passive or active levels of management throughout data collection. Christy (2014) described an intervention where interviewers were provided the estimated response propensity of cases in their workload to help interviewers more efficiently plan their contact attempt strategies. Alternatively, an interviewer whose cases are approaching (or exceeding) expected lag times might prompt feedback from a supervisor, and the earlier that intervention can occur, the faster an improvement might be made (Edwards et al. 2020). Walsh and Coombs (2013) described a reassignment strategy based on a different parameter, the scaled evenness of finding attempts (SEFA), which they find was related to higher rates of contact and completion. To order to intervene during a data collection period, as the examples above suggest, the indicators or parameters of interest must be generated during the data collection period itself. However, the first step to being able to intervene to reduce these time lags is to estimate their expected length.

Other survey data collection parameters related to data quality, survey progress, and survey cost are commonly generated and monitored as part of data collection operations. These parameters may be defined at various levels of the data collection process, including the interviewer level (Mneimneh et al. 2018); the question-level (Mittereder and West 2021); the case-level (Wagner et al. 2012), and the survey level (LaFlamme and Karaganis 2010). Indicators of cost, progress, and quality have also served as inputs into responsive and adaptive designs (Groves and Heeringa 2006; Schouten et al. 2011) that allow for the balancing of cost and quality goals through centralized interventions including mode changes (Chesnut 2013; Coffey et al. 2019), case prioritization (Walejko and Wagner 2018; Tolliver et al. 2019; Peytchev et al. 2020) and stopping rules (Wagner and Ragunathan 2010; Peytchev 2014).

Of the varied data collection parameters that can be predicted, case-level response propensity is the most commonly estimated and used for interventions. Many implementations of response propensity prediction during survey data collection have used only the data available from the in-progress collection period to estimate coefficients in a logistic regression or probit model (Groves and Heeringa 2006; West and Groves 2013; Schouten et al. 2009, 2011; Chesnut 2013; Coffey et al. 2019). The data sources available for estimating response propensity in this setting include survey frame and sampling data and paradata (Couper 2000, 2017) for the in-progress period of data collection as it accumulates after each operation or contact attempt.

Other examples use only data from the past implementations of a survey to estimate coefficients in the response propensity model, which are then applied to covariates during the in-progress data collection period (Peytchev et al. 2010; Roberts et al. 2014; Calinescu et al. 2013; Jackson et al. 2020). Unlike the prior examples, these applications do not consider information learned during the in-progress data collection period.

More recent applications have incorporated both historical data and data from the in-progress data collection period in a Bayesian framework to improve predictions of response propensities (Schouten et al. 2018; West et al. 2021). Both studies found that Bayesian methods offered modest improvements in predictions of daily response propensities compared to predictions relying on either historical or current data alone, particularly when model misspecification was small.

For this application, we propose a model to predict our parameter of interest, the time lag between the first contact attempt and the first respondent contact in a face-to-face interviewer-administered survey. We use three sets of data for predicting lag: historical collection period data in a non-Bayesian prediction; current data from the in-progress data collection period data in a non-Bayesian prediction; and combining historical data and current data in a Bayesian prediction setting. The purpose of this research is to understand whether the Bayesian framework that led to improvements in predictions of response propensity can also lead to similar improvements in predictions of a time-to-event parameter, the estimated day lag between first contact attempt and first respondent contact. We apply our methodology to predict these contact lags in the National Health Interview Survey (NHIS), using monthly survey data from July 2014 through June 2016. We compare the case-level error in prediction of lags when using the Bayesian framework to methods that leverage data from either historical or in-progress data collection periods using maximum likelihood estimation. We also vary the strength of the priors put on the coefficients in the Bayesian framework to better understand how the strength of the priors affects prediction error. Additionally, we use the predictions generated from the different methods to illustrate how the different methods would perform when trying to identify cases with short or long lags between the first contact attempt and first respondent contact. Our goal is to contribute to the growing literature on the use of Bayesian methods for improving predictions of survey data collection parameters. After introducing the NHIS survey data (Section 2) we introduce a Weibull hurdle model which is used to predict the lag between first attempt and first contact and describe our method for prior construction (Section 3). We then present our analytic approach and evaluation methods (Section 4). Results are presented (Section 5) followed by a discussion (Section 6).

2. Description of Data

2.1. *The National Health Interview Survey*

The National Health Interview Survey (NHIS) is a monthly cross-sectional household interview survey that serves as the principal source of information on the health of the civilian noninstitutionalized population of the United States and is one of the major data collection programs of the National Center for Health Statistics (NCHS 2018).

The sample design follows a multistage area probability design that permits the representative sampling of households and non-institutional group quarters (e.g., college dormitories). The entire United States is divided into approximately 1,700 primary sampling units (PSUs), which can consist of a county, a small group of adjacent counties, or a metropolitan statistical area. PSUs are selected, some with certainty (self-representing PSUs), and others with probability proportional to size (non-self-representing PSUs). Within those selected PSUs, clusters of households and non-institutional group quarters are selected. Additionally, an oversample is taken for geographies with higher expected populations of particular age and race/ethnicity groups. Interviewing for the NHIS is conducted continuously throughout each calendar year.

The U.S. Census Bureau is the data collection agent for the NHIS. Survey data are collected continuously throughout the year by Census interviewers. Cases released into the field are attempted for up to 30 days. The NHIS consists primarily of face-to-face interviews conducted in respondents' homes, but follow-ups to complete interviews may be conducted over the telephone. A telephone interview may also be conducted when the respondent requests a one or when road conditions or travel distances would make it difficult to schedule a personal visit before the required completion date (NCHS 2018). We used two years' worth of monthly data collection periods, spanning the months from July 2014 to June 2016 for this work.

2.1.1. Auxiliary Data Sources

Four data sources were used for this evaluation, in addition to the NHIS sample itself. The Census Bureau Planning Database (PDB) is a detailed dataset including sociodemographic information at the block group level that is produced annually using data from the five-year ACS estimates (Census 2008) and the Decennial Census. For this application, we used the 2016 release. For a full list of items on the PDB, see Census (2016). Second, we obtained a dataset of basic employment information about interviewers, including which regional office an interviewer belongs to and their experience level on the NHIS.

We also utilized two sources of paradata, the Neighborhood Observation Instrument (NOI) and the Contact History Instrument (CHI). In the NOI, interviewers are asked to record information about the housing unit and neighborhood from their own observations. These neighborhood observations require no contact with the respondent and should be recorded just prior to the first contact attempt. For the CHI, on the other hand, interviewers record the date, time and outcome of each contact attempt and information about interactions with sample persons. Additionally, the CHI includes information about the field management structure, including which interviewers are assigned to work each case and whether a case was reassigned during the field period. These data sources were selected for this application partially because of their availability, but also because they

Table 1. Auxiliary data sources, example data items and their availability for use in prediction of lag.

Data source	Available prior to start of the data collection period?	Fixed or time-varying	Example items
NHIS Sample File	Yes	Fixed	Sample unit case ID Census block-group-level ID Data collection sample month Census fieldwork regional office
Planning Database	Yes	Fixed	Census block group (BG) ID % HU in a BG that are mobile homes % Population in a BG that is urbanized % Pop in a BG with a college degree
Interviewer Information	Yes	Fixed	Interviewer ID New NHIS interviewer (< one year experience)
Interviewer Observations	No	Fixed	Sample unit case ID Day/time of contact attempt Interviewer ID Evidence of children at the housing unit (HU) Evidence of smoking at the HU Bars observed on windows of the HU
Contact history Instrument	No	Time-varying	Sample unit case ID Interviewer ID Day/time of contact attempt Outcome of contact attempt Activities completed on contact attempt Case reassignment indicator

include the types of data identified in Groves and Couper (1998) for predicting outcomes like household contactability, survey cooperation, and response propensity.

Table 1 lists the five data sources, whether the information contained in each source would be available prior to the start of data collection, whether the data items from the source are fixed or time-varying during the data collection period, and some example items from each source. For a full list of items included in the predictive models as part of this research, see the online supplementary materials.

Identifiers across the different data sources can be used to link the sources and generate a file at the contact attempt level for each sample unit in each data collection month. It is important to note that, until the interviewer makes their first contact attempt at a sample unit, the information from the NOI and the CHI is missing. Both of those data sources are only reported at or after the point-in-time of the first contact attempt, which restricts when they are available for prediction of the lag between first attempt and first contact.

2.2. *Parameter of Interest: Lag Between First Contact Attempt and First Respondent Contact*

For this study, we analyzed data from 24 months of the NHIS, ranging from July, 2014 through June, 2016. Our parameter of interest was the time lag, in days, between the first contact attempt made on a sample unit, and the first respondent contact. We calculated this parameter at the case-level from the CHI data by subtracting the data collection day of the first attempt from the data collection day when the first contact with the respondent was reported. If respondent contact was made on the first contact attempt, the resulting lag was zero days. Otherwise, all lags are positive integers. Table 2 summarizes, by month, the approximate sample size in housing units of the NHIS, the percent of cases with a lag greater than zero, the mean and variance of the lags conditional on the lag being greater than zero, and the annual response rate for the NHIS. While the NHIS is conducted monthly, response rates are only reported yearly (NCHS 2018). Table 2 shows that the percentage of cases with a positive lag, and the means and variances of non-zero lags, are relatively consistent across the 24 months of data.

While the mean lag across months was relatively similar, the mean lag *during* any one month varies by the day of the data collection month when the first attempt is made. Again, the day of the first contact attempt was taken from the CHI data. Table 3 shows that, for example, cases first attempted during the first week of data collection have a mean lag of nearly nine days, while cases first attempted within the last week of the month have a mean lag of less than four days. This suggests that the day of the first attempt may be an

Table 2. Mean and variance of lag in days between 1st attempt and 1st contact by interview month.

Interview period	Number of cases	% Cases with lag > 0	Mean(lag) if lag > 0	Var(lag) if lag > 0	Response late (AAPOR RR6)
July 2014	5200	0.4171	8.890	48.67	73.8% annualized
Aug 2014	5100	0.3990	8.216	41.90	
Sept 2014	4900	0.4131	7.941	42.54	
Oct 2014	5300	0.4236	8.009	43.89	
Nov 2014	5000	0.3895	7.837	40.78	
Dec 2014	5100	0.4058	9.192	48.30	
Jan-2015	5600	0.3861	8.178	39.93	70.1% annualized
Feb 2015	5400	0.4181	8.464	44.49	
Mar 2015	5400	0.3984	8.658	49.41	
Apr 2015	5800	0.4165	8.396	41.84	
May 2015	5400	0.4026	8.424	45.74	
Jun 2015	5600	0.3842	8.072	43.88	
Jul 2015	5400	0.3938	8.589	40.90	
Aug 2015	5200	0.3929	8.385	46.09	
Sept 2015	5100	0.4003	8.449	43.12	
Oct 2015	5300	0.4015	8.847	49.48	
Nov 2015	5000	0.3754	8.062	42.60	
Dec 2015	4800	0.3915	9.684	59.51	
Jan-2016	5200	0.3876	7.797	35.65	67.9% annualized
Feb 2016	6000	0.3952	8.657	42.47	
Mar 2016	6600	0.3893	8.861	51.00	
Apr 2016	6000	0.3947	8.665	43.58	
May 2016	5700	0.3951	8.448	44.40	
Jun 2016	5800	0.4009	8.724	47.89	

Table 3. Mean and variance of lag in days between 1st attempt and 1st contact by week of 1st attempt, all months.

Week of data collection	Mean(lag) if lag > 0	Var(lag) if lag > 0
1	8.891	48.81
2	7.763	33.56
3	5.770	16.94
4	3.716	9.944

important predictor of the lag between first attempt and first contact. Lags in Table 3 were averaged over all 24 months of data collection used in our application.

These results make intuitive sense. As a fixed data collection period progresses, there are fewer remaining data collection days for cases to be contacted, and so interviewers may make contact attempts closer together, or at higher frequency to make contact with the sample member. Additionally, interviewers may have fewer remaining cases in their workloads, and so more attention can be paid to those remaining cases.

3. Modeling the Day Lag Between First Contact Attempt and First Respondent Contact

In order to model the day lag between first contact attempt and the first respondent contact, we needed to identify a model that can properly account for a lag of zero due to a contact being made on the first attempt. Additionally, our model needed to account for covariates that are available prior to the start of data collection as well as covariates available only after the first contact attempt is completed. As a result, we employed a hurdle model to combine two different processes – the likelihood of making contact on the first attempt, and a time-to-event, or survival, model to predict the length of the lag between the first attempt and first contact, given that contact was not made on the first attempt.

3.1. Hurdle Models

A hurdle model (Mullahy 1986) is useful when one believes there are two separate processes at work – one that generates zeros (in our case, "zero lags", where the interviewer makes contact on the first attempt) and the other that generates some non-zero value, in our case a positive lag (Ma et al. 2015). Hurdle models are similar to zero-inflated models (Rose et al. 2006) in the fact that they are useful for handling data with excess zeros. However, zeros are conceptualized differently between the two classes of models. Hurdle models are a mixture model where one component predicts whether a zero or non-zero status (the "hurdle"), and the second component estimates a positive value conditional the non-zero status. A zero-inflated model, on the other hand, allows zeros to appear in either component of the model. That is, even if a non-zero status is predicted, there is the chance that the second component of the model still results in a zero. Here, the hurdle model was conceptually intuitive. If the first contact attempt results in a noncontact, we would expect a nonzero lag for the case to result. The negative binomial model construction would suggest that when there is a noncontact on the first attempt, there could still be a zero lag. See Feng (2021) for more discussion of zero-inflated and hurdle models.

We used a logistic regression model to predict whether a noncontact would result on the first contact attempt, resulting in a non-zero lag. The covariates available for this portion of the hurdle model were restricted to those known prior to the first contact attempt, which are found in the sample file and the PDB, as shown in Table 1. For the second component of our model, which estimates the length of the non-zero lag in cases that have a first contact, we chose to use a survival model to allow us to retain right-censored cases in our analytic data sets. This means that cases for which the first contact attempt resulted in a noncontact, and for which contact with a respondent had not been made, could still be included in the set of cases used to estimate model coefficients. While only a small portion of the cases (less than 2% in any given month) were right-censored at the end of data collection, the use of a survival model is particularly salient when considering prediction methods using data from the in-progress data collection period, where many of the cases that will ultimately be contacted after nonzero lags have not yet had attempts that result in contact. If we chose an alternative count model, such as a negative binomial, only cases that had both a noncontact on the first attempt *and* an observed first contact would be included in the set of cases used to estimate model coefficients. For more information on how censored cases are still included in parameter estimates of survival models, see the online supplementary materials.

To determine which parametric survival distribution best fit our data, we compared the empirical distribution of positive lags (without controlling for any factors) in the NHIS data against Poisson, Weibull, Gamma, and Lognormal distributions both visually and with goodness-of-fit statistics, using the `fitdistrplus` package in R. Figure 1 below shows panels for each of the four distributions evaluated for one data collection period to illustrate how different time-to-event distributions compare with the empirical distribution for a given month.

The black bars and density curve display the empirical distribution of the actual lag between first attempt and first contact, based on the NHIS data, while the red density curve

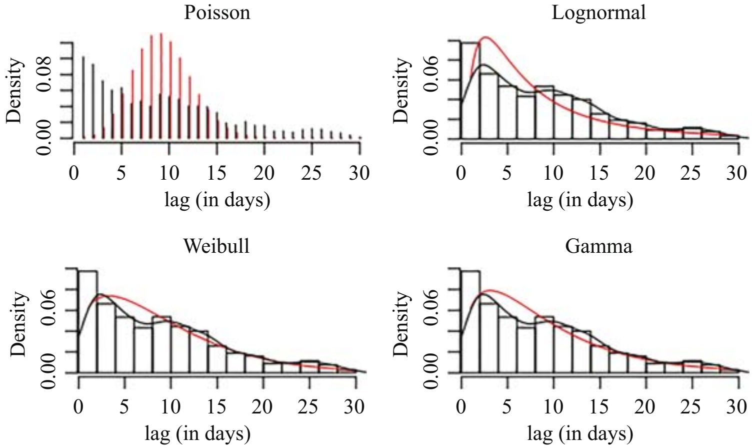


Fig. 1. Four parametric survival distributions compared to empirical distribution of lag (in days).

(or histogram, in the case of the Poisson distribution) represents the theoretical distribution when parameters are estimated from the NHIS data. The Poisson distribution provided the worst fit, followed by the lognormal distribution. The Weibull and Gamma distributions were similar in their fits, and so to decide, we examined Chi-Square tests, AIC, and BIC. For the month displayed in Figure 1, and across the 24 months of our data, the AIC and BIC scores of the Weibull distribution were smaller, and the Chi-Square test was larger, more often. As a result, we elected to use a Weibull distribution for model fitting and prediction.

3.2. The Weibull Hurdle Model

In this setting, the prediction of interest, y_i, is the length of time in days (lag) that will elapse between the first contact attempt and the first contact with a household sample member for the i^{th} case. For each observation, with probability π_i, $y_i = 0$, and with probability $(1-\pi_i)$, $y_i > 0$. Thus the probability distribution function for an observation can be written as:

$$f\left(y_i;\beta,\,\sigma,\gamma\right) = \begin{cases} 0 & \textit{with probability}\, \pi_i = \,\exp\left(\gamma' z_i\right)/(1+\,\exp\left(\gamma' z_i\right) \\ g\left(y_i,\beta,\,\sigma\right) & \textit{with probability}\,(1-\pi_i) = 1/\left(\left(1+\,\exp\left(\gamma' z_i\right)\right)\right. \end{cases}$$

where $g(y_i, \beta,\, \sigma)$ is the probability distribution function (p.d.f.) of the Weibull distribution and z_i are the covariates that govern the log-odds of the probability of a zero-lag contact. The likelihood function for the Weibull model in the presence of right censoring is:

$$L = \prod_{i=1}^{n} S(t_i)(\lambda(t_i))^{\delta_i},$$

where $S(t_i) = 1 - G(t_i;\ \beta,\ \sigma) = \ \exp(-\,exp\,(x_i^{'}\beta/\sigma)t_i^{1/\sigma})$ is the survival function, $\lambda(t_i) = \frac{1}{\sigma}\,exp\,(x_i^{'}\beta/\sigma)t_i^{((1/\sigma)-1)}$ is the hazard function, δ_i is the censoring indicator (equal to 1 if the failure time is observed and 0 if not), $t_i = \ \min(y_i, c_i)$ is the observed failure or censoring time, and x_i are the covariates that govern the log of the number of days until a non-zero lag contact. The full likelihood can then be written as follows:

$$L\left(\beta,\sigma,\gamma\right) = \prod_{i=1}^{n}\left[\left(\frac{exp\,(\gamma' z_i)}{1+\,exp\,(\gamma' z_i)}\right)^{1-d_i}\left(\left(\frac{1}{1+\,exp\,(\gamma' z_i)}\right)exp\left(-\,exp\left(\frac{x_i{}'\beta}{\sigma}\right)t_i^{\frac{1}{\sigma}}\right)\right.\right.$$
$$\left.\left.\left(\frac{1}{\sigma}\,exp\left(\frac{x_i{}'\beta}{\sigma}\right)t_i^{((1/\sigma)-1)}\right)^{\delta_i}\right)^{d_i}\right]$$

where d_i is an indicator specifying there *will be* a lag between the first attempt and first contact and now z_i are the covariates that govern the probability of a zero lag (note that x_i and z_i can be disjoint, partially overlapping, or equal to each other). The first term in the likelihood represents the probability that there will *not* be a lag between the first event and the first contact; that is, contact will be made on the first attempt. The second term represents the probability that there will be a positive lag, and the last term in the likelihood represents the survival portion of the model, accounting for censoring. (Note that there are alternative parameterizations of the Weibull distribution that, e.g, replace $x_i{'}\beta$ with $\alpha_i = \exp\left(x_i{'}\beta\right)$ [scale parameter] and $\beta = \frac{1}{\sigma}$ [shape parameter].)

The likelihood includes parameters for both the binary process and the time-to-event process. Maximum likelihood estimation shows that the different contributing terms can be maximized separately allowing the model to be estimated in parts, computationally. Specifically, the log-likelihood can be written as:

$$l\left(\beta,\sigma,\gamma\right) = \sum_{i=1}^{n}\left[(1-d_i)\gamma' z_i - \mathit{log}\,(1 + \mathit{exp}\,(\gamma' z_i)\right]$$
$$+ \sum_{i=1}^{n} d_i\left[\left(-\,\mathit{exp}\left(\frac{x_i{'}\beta}{\sigma}\right) t_i^{\,1/\sigma}\right) + \delta_i\left(\frac{x_i{'}\beta}{\sigma} + \left(\frac{1}{\sigma} - 1\right)\mathit{log}\,t_i - \mathit{log}\,\sigma\right)\right]$$

As a result of this construction, the portion of the model that predicts the binary outcome (lag/no lag) can be estimated using logistic regression from all observations, and the portion of the model that predicts the lag length, can be estimated using Weibull regression among only those cases with a non-zero lag (Smithson and Merkle 2013, chap. 5). However, we refer to this as a single model, as the goal of this prediction is to determine the expected lag between first attempt and first contact, and both portions of this model are needed to arrive at that prediction.

3.3. Prior Derivation for Bayesian Prediction

We derive prior distributions for the coefficients in our Weibull hurdle model with the goal of improving the accuracy of predictions of those coefficients, in order to then improve the predictions of lag between first attempt and first contact during the data collection period. We assume that historical data will provide informative priors for each of our model coefficients. We accomplish this by first fitting a Weibull hurdle model using three months of historical data to obtain parameter estimates $\left(\hat{\gamma},\hat{\beta},\hat{\sigma}\right)$, and the associated estimated variances $\hat{V}\left(\hat{\gamma},\hat{\beta},\hat{\sigma}\right)$. These parameters capture the time-varying nature of some of the covariates within a single data collection period in a survival model framework.

Assuming approximate normality by the properties of maximum likelihood estimates, we form priors, $p\left(\gamma,\beta,\sigma\right) = N\left(\left(\hat{\gamma},\hat{\beta},\hat{\sigma}\right), c\hat{V}\left(\hat{\gamma},\hat{\beta},\hat{\sigma}\right)\right)$, where c is a constant that controls the degree to which the prior information is used in the daily estimation procedure. (A normally distributed prior distribution is a logical choice given the asymptotic normality of maximum likelihood estimates.) We assume that the estimates of the model coefficients will have similar precision each month, due to similar sample sizes. As a result, we would expect that the prior estimates for the coefficients, which are based on three months of data instead of just one, would be three times as precise as the coefficient estimates generated from just the in-progress data collection period. As part of a sensitivity analysis, we inflate the variance of each of the priors by a factor of c. By varying the inflation factor as 1/3, 1/2, 1, 2, and 3, we varied the precision of the prior coefficients to represent 27 months, 12 months, 3 months, 3/4th of a month, and 1/3rd of a month, respectively.

3.4. Implementation

The posterior distributions for coefficients of all covariates in the Weibull-hurdle model are estimated each day using the priors and all data through the most recent day from the in-progress data collection period. The `brms` package was then used to conduct resampling and estimation for Bayesian inference of parameters and estimation of posterior predictions. The code for generating predictions via all methods for a given day and month are provided in the online supplementary materials.

4. Methods

4.1. Predicting Day Lag to First Contact: Four Methods

For this application, we used 24 months of data from the NHIS, covering the time period from July 2014 through June 2016. For each prediction data collection period, the prediction of interest was the length of time in days that would elapse between the first contact attempt and the first contact with a household sample member. In addition to the *Bayesian method* described in Section 3, we consider three alternative approaches for comparison:

1. The first method uses only available data from the in-progress data collection period to estimate coefficients for the covariates in a Weibull hurdle model, and then uses those coefficients to predict the expected lag for each open case in the current month. We refer to this as the *current method.*
2. The second method uses only historical data to estimate the mean expected lag. The expected lag for all cases in the current month, then, is just the overall average lag of the three prior months, ignoring any additional information. This leads to all sample units having the same predicted lag in a given month. We refer to this as the *mean method.*
3. The third method also uses historical data but estimates coefficients for the covariates in a Weibull hurdle model, and then uses the point estimates of those coefficients to predict the expected lag for each open case in in the current month. We refer to this as the *historical method.*

For the *mean, historical,* and *Bayesian* methods, which all leverage historical data, three consecutive months were used as the historical data, and the next month was considered the prediction period of interest. For example, if October 2014 was the predictive period of interest, July, August and September of 2014 would be used to generate the mean for the *mean* method, coefficients and standard errors for model covariates for the *historical* method, and point estimates and standard errors for use as priors in the *Bayesian* method. Table 4 summarizes which months of data are used for each method for a prediction period of interest.

We generate predictions at the case- and day- level for each of the 21 time periods of interest (October 2014–June 2016), giving us a set of daily predictions for each case that has not yet had a successful contact.

4.2. Evaluation Measures

The four methods will be compared primarily using measures of mean prediction bias, root mean squared error, and percent change between any two of the four methods. Cases in the in-progress month for which we are evaluating our predictions are included in these calculations if no contact with a sample member has been made on the day the prediction

Table 4. Summary of prediction methods by data types used.

Prediction method	Months used to estimate model coefficients	Coefficient parameters	Prediction period of interest
Current	October 2014 (in-progress)	Coefficient point estimates	October 2014
Mean	July 2014 August 2014 September 2014	Mean predicted lag	
Historical	July 2014 August 2014 September 2014	Point estimates for coefficients in Weibull hurdle model	
Bayesian	*Priors:* July 2014 August 2014 September 2014 (priors) *Likelihood:* October 2014 (in-progress)	Point estimates, standard errors for coefficients in Weibull hurdle model based on historical data; Posterior estimates and standard errors for coefficients in Weibull hurdle model based on priors plus October 2014 accumulated data	

has been made, and a first contact attempt was made prior to the day the prediction was made. Mean prediction bias (MPB) is expressed as:

$$MPB^m_{d_j} = \frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{y}^m{}_{id_j} - y_i\right)$$

and root mean square prediction error (RMSE) is expressed as:

$$RMSE^m_{d_j} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n}\left(\hat{y}^m{}_{id_j} - y_i\right)^2}$$

where n is the number of cases eligible for inclusion in the error estimate (this varies throughout the prediction period of interest), $\hat{y}_i$ is the predicted value of the lag, y, for the i^{th} case under the m $^{\text{th}}$ method on the d $^{\text{th}}$ day, for cases first attempted on day j, and y_i is the true value of the lag, y.

The percent change between the RMSE for the m^{th} method and the method selected as the baseline, b, on the d $^{\text{th}}$ day for cases first attempted on day j is calculated as:

$$PCHG^{m_1 m_b}_{d_j} = 100{*}\left(\frac{RMSE^m_{d_j} - RMSE^b_{d_j}}{RMSE^b_{d_j}}\right)$$

Percent change in the MPB can be calculated similarly. Creating measures of MPB, RMSE, and PCHG based on d_j, the cutpoint by day the case was first attempted leads to a different number of summary error estimates on each day. For example, on day $d = 4$, for each of the methods, there would be three separate estimates of MPB, one each for $j = 1, 2, 3$, where j denotes the day the case was first attempted. One day $d = 25$, there would be up to 24 separate estimates for MPB. Days where no cases were attempted for the first time would not generate an estimate for MPB. This allows us to not only examine the distribution of MPB and RMSE at each cut point, but also examine how the distribution of error in predicted lag differs by how close to the day of first attempt the predictions are made.

The expectation is that the *Bayesian* method will produce predictions of the expected lag closer to that of the actual lag, resulting in a smaller overall MPB and RMSE, as the modeling procedure is effectively borrowing strength across the historical data and current data to make a prediction. For this evaluation, we do not use design-adjusted variance estimates.

4.3. Evaluation Methods

In order to compare the four discussed methods for predicting lag (*current (C), mean (M), historical (H)*, and *Bayesian (B)*), we first compare the MPB and RMSE of the prediction of the lag for each of the four methods using boxplots that illustrate the mean and range of the errors in prediction of lag for eligible cases on each day. Predictions of the expected lag and the resulting MPB and RMSE of those predictions depend not only on the prediction method used, but also on when the prediction is made during data collection, which we refer to as the cut point. This is because each day, new cases are attempted for the first time, and contact is being made in other cases. Therefore, we are attempting to predict the expected lag for a pool of cases that is changing on a daily basis. In order to evaluate the quality of predictions, we evaluate MPB and RMSE at several cut points in data collection, after days 2, 4, 6, 8, 10, 15, 20 and 25.

Additionally, we are particularly interested in the quality of predictions of expected lag shortly after the first contact attempt is made. If we are able to correctly predict the lag close to the day of the first contact attempt, we maximize the window where we can implement an intervention to inform the interviewer or even try to reduce that expected lag. To see this, we recreate the initial plots comparing MPB and RMSE, but restrict them to cases who have had their first contact attempt occur within two days of the cut point. In other words, on day 6, these boxplots would only display the mean and range of errors in prediction of lag for cases first attempted on day 4 or 5. On day 10, these plots would be limited to cases first attempted on days 8 or 9.

Additionally, in the online supplementary materials, we include an example of how this prediction could be used during a data collection period to identify cases that are likely to have a long lag.

5. Results

5.1. Comparison of Four Predictive Methods (MPB and RMSE)

For each cut point during each of the 21 time periods, we used all open, uncontacted cases to generate estimates of MPB and RMSE. The boxplots were then generated using the $MPB^m_{d_j}$ and $MPB^m_{d_j}$ for all m time periods.

Figure 2 displays estimates of MPB for predictions of the lag and shows that all methods underestimate the actual lag between first attempt and first contact throughout the data collection process. Additionally, as data collection goes on, the underestimation MPB increases. It is also evident that, until day 15, all methods that use historical data in some form (*M, H, and B)* outperform the c*urrent* method with respect to prediction MPB. While the three methods that use historical data all perform similarly late in data collection, the *historical* and *Bayesian* methods perform better than the *mean* method until day 10. The *Bayesian* method appears to provide small improvements over the *historical* method over the 21 time periods, as evidenced by central tendencies of estimates of the MPB being closer to zero. The use of historical data is particularly helpful in reducing bias in the very early period, privileging the stronger priors; by day 4 however, the prior has little impact.

In addition to the overall predictive abilities of the four different methods, we are particularly interested in the temporal effectiveness of the predictions. In other words, we are interested in how well we can predict the expected lag near the day of the first attempt, when we are close to the entry point of a given case into the dataset. Just after the initial contact attempt, we have the least data and the most time to implement an intervention if needed. In order to explore this, we generated Figure 3, which is equivalent to Figure 2 but restricted the to include only measures of error based on cases worked within two days of the cut point. So, when $d = 4$, only cases that were first attempted on day 2 or 3 are included in the estimates of MPB. On day 25, only those cases first attempted on days 23 and 24 are included.

While Figure 3 follows the same general pattern as Figure 2, there are some differences. Most notable is that as time progresses through the data collection period, the MPB in predicting lag for recent cases is much smaller (the bias is closer to zero) for all cases, whereas in Figure 2, the under-estimation bias increases over time. This finding shows that the model is relatively accurate at predicting lag length close to the day of the first attempt, whether that lag is long or short. However, as time moves away from the initial contact attempt, the quality of predictions of the lag decreases for the outstanding cases. This suggests that cases where we underestimate lag, we underestimate it by a large amount. Again, the *current* data method performs the worst until late in data collection, but here, it never really outperforms the *historical* or *Bayesian* methods. Late in data collection, the *mean* method begins outperforming other methods, but this is not consistent throughout data collection. Again, we see that the central tendencies of the MPB in the *Bayesian* method are closer to zero than in the *historical* method, demonstrating a small improvement. Among the *Bayesian* methods, we again see an advantage for the strong prior at day 2, but little difference from day 4 on.

We see the same pattern in the RMSE of the lag predictions in Figure 4. The *current* method performs worse than the other methods until mid-way through the data collection period, and the other three methods are competitive with each other. Again, the *Bayesian* method appears to provide improvements over the *historical* method until late in the data collection period, as the central tendencies and intraquartile ranges are smaller in the *Bayesian* method than in the *historical* method. As with bias, there is a tendency for the stronger prior to perform better with respect to RMSE at day 2, but then little difference thereafter.

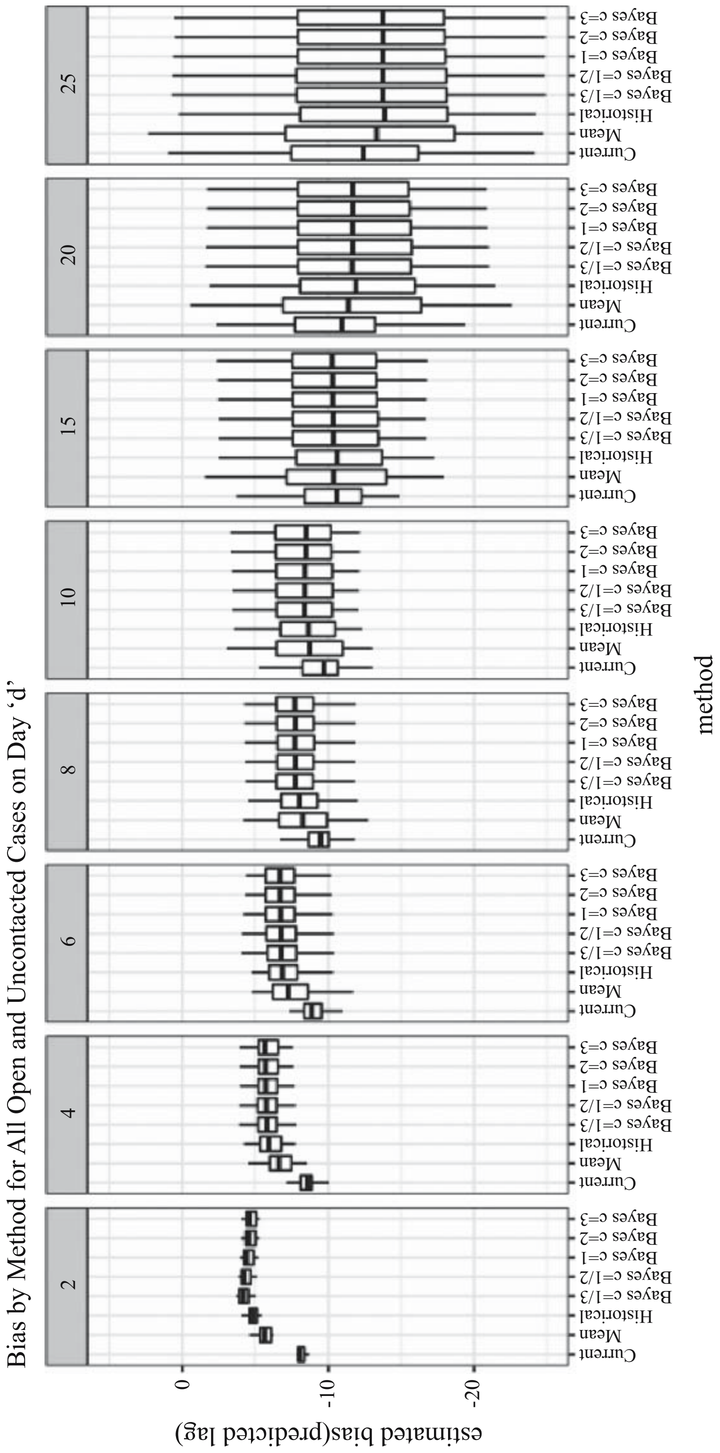


Fig. 2. Bias by prediction method for all open and uncontacted cases on cut point day "d".

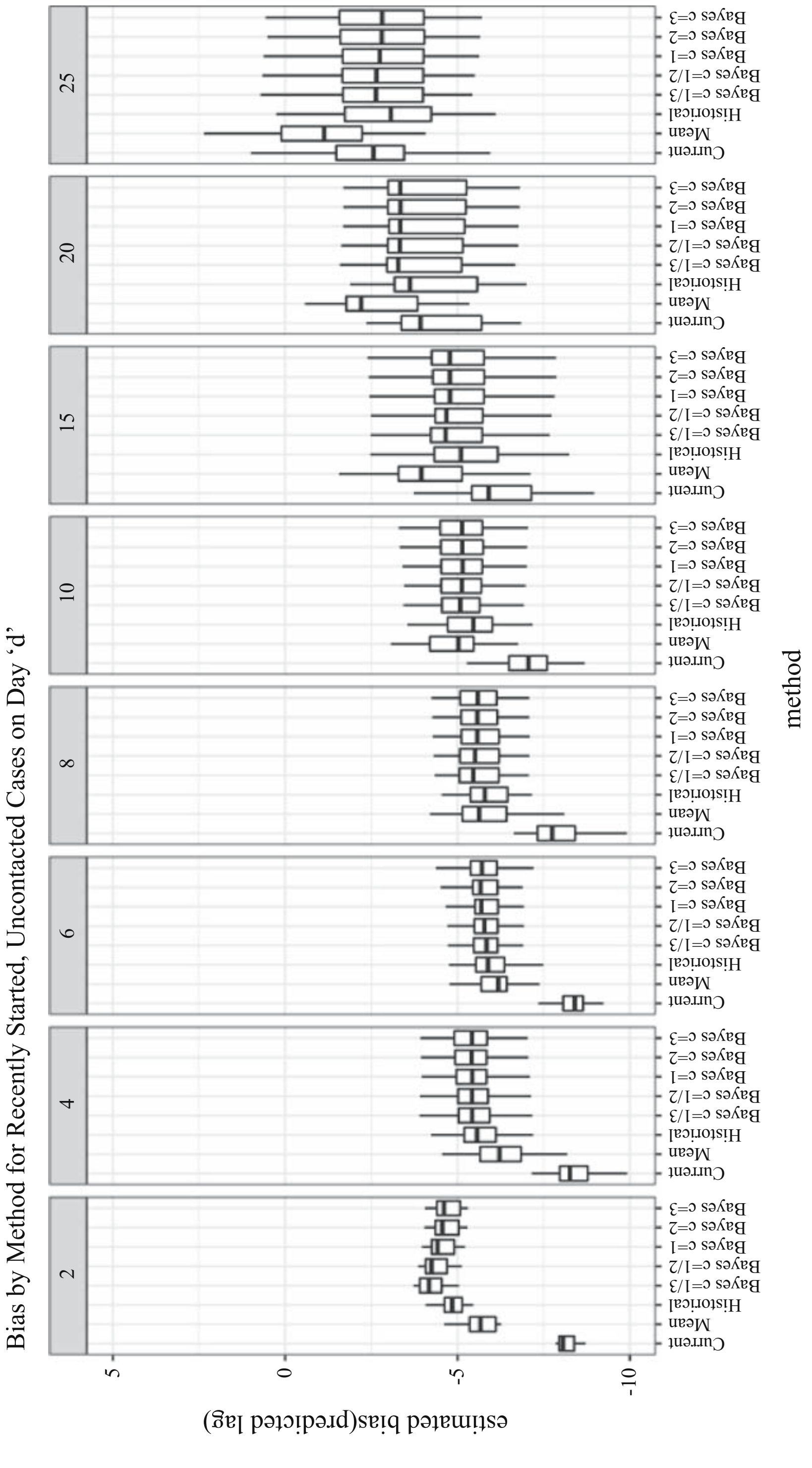


Fig. 3. Bias by prediction method for recent open and uncontacted cases on cut point day "d".

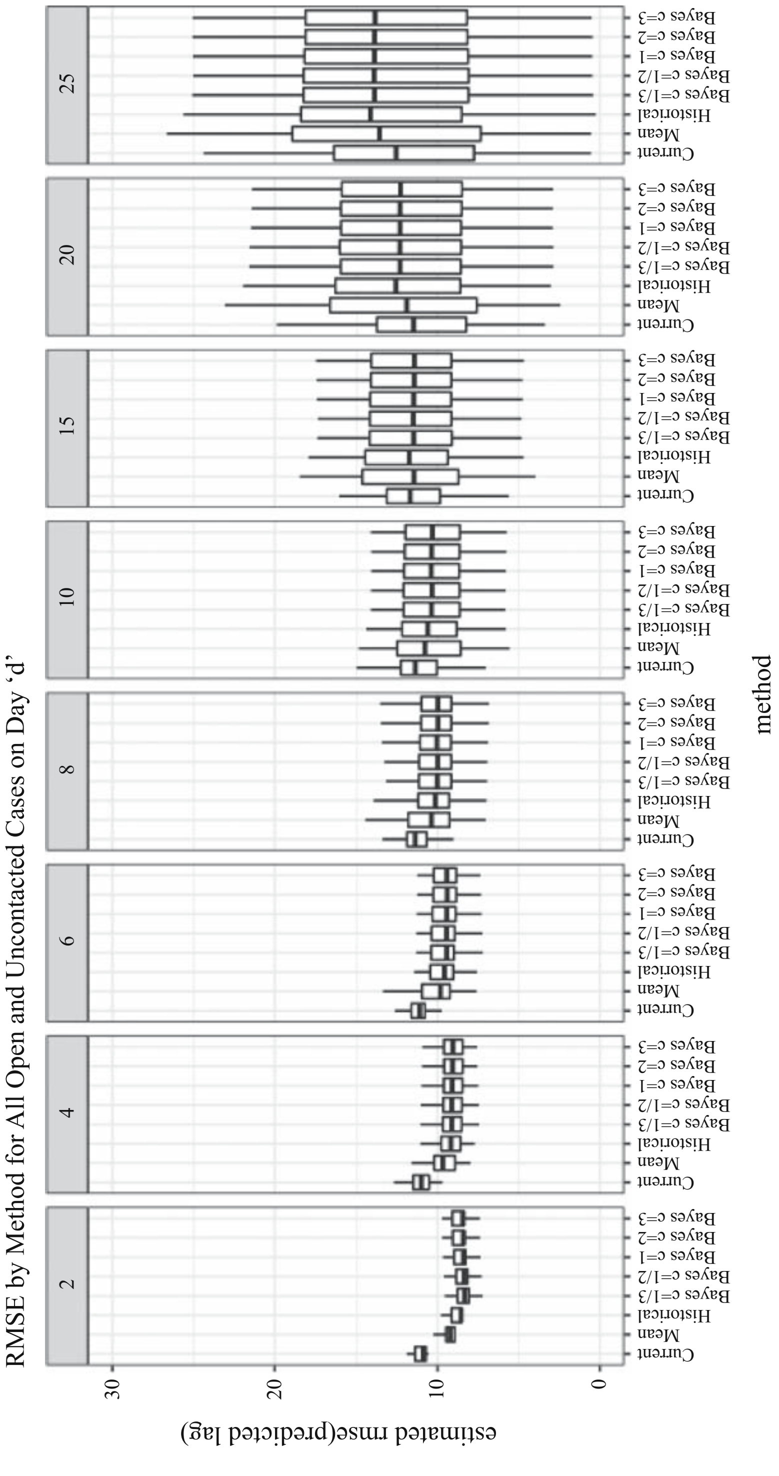


Fig. 4. RMSE by prediction method for all open and uncontacted cases on cut point day “d”.

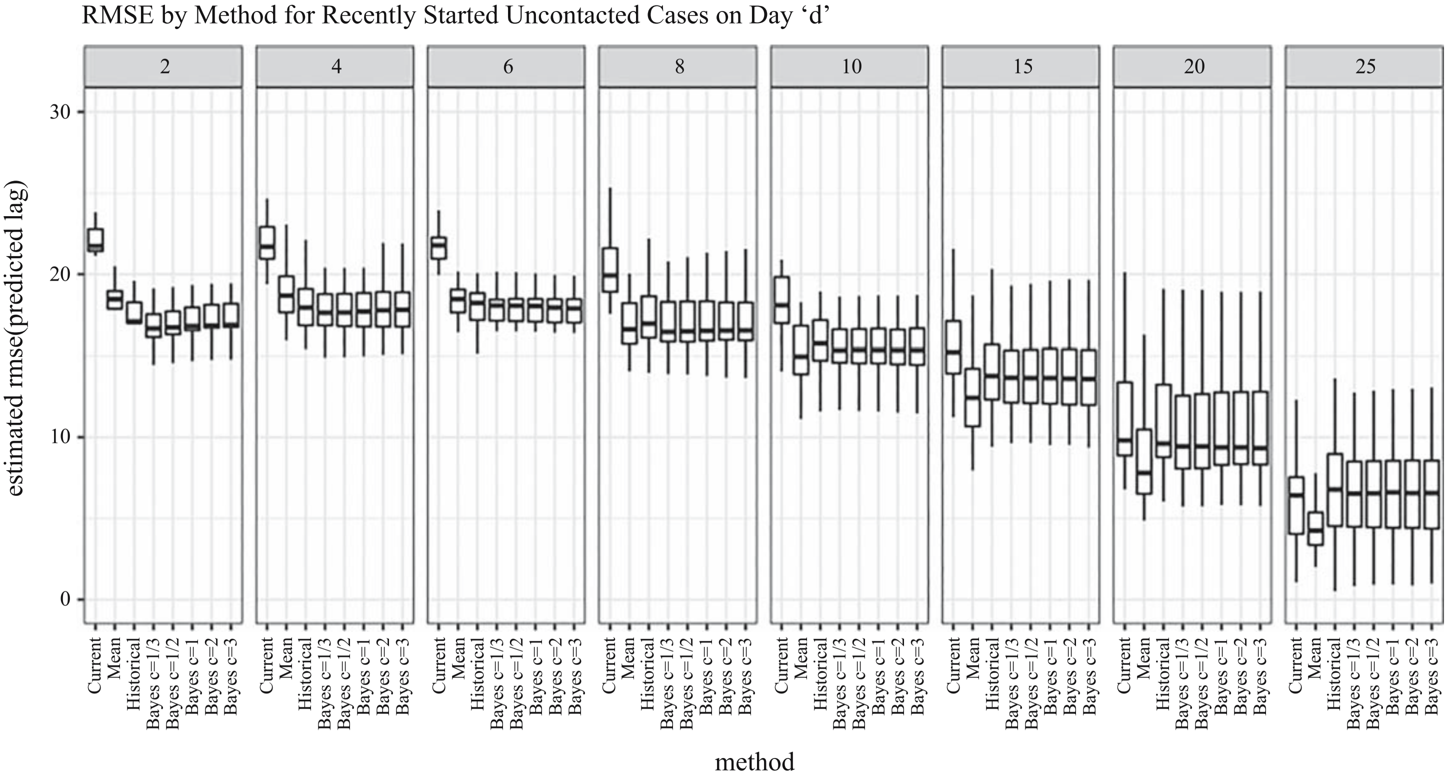


Fig. 5. RMSE by prediction method for recent open and uncontacted cases on cut point Day “d”.

Figure 5 displays the RMSE of predictions of lag for recent cases and continues to demonstrate that the use of historical information is generally helpful for improving predictions. Additionally, throughout data collection, the *Bayesian* method provides small improvements in the central tendencies of RMSE over the *historical* method. Again. there is a tendency for the stronger prior to do better with RMSE at Day 2, but then little difference thereafter.

Figure 6 plots the percent change in RMSE and MPB of the *Bayesian* method where the prior is equivalent to 1/3 of a month from the *historical* method. Each point on the scatter plot represents a cut point (prediction day) by day of first attempt. Cases receiving their first contact attempt in the prior two days and cases attempted more than two days prior to the cut point are plotted separately. For example, one data point in the plot of recent cases would be for the measures of MPB and RMSE on Day 4, for cases first attempted on Day 2, in a given data collection period.

Most of the data points fall in the southwest quadrant, representing a reduction of both RMSE and MPB in the *Bayesian* method when compared to the *historical* method. This is true for both recent and older cases. Additionally, when there are improvements in MPB and RMSE, those improvements have a larger range (reductions in the bias and RMSE reaching 20% and 10% respectively) than the situations where the bias or RMSE increase when the *Bayesian* method is used (increases are generally limited to 10% increases in MPB and 5% increases in RMSE). These results suggest the *Bayesian* approach can provide modest improvements to our predictions.

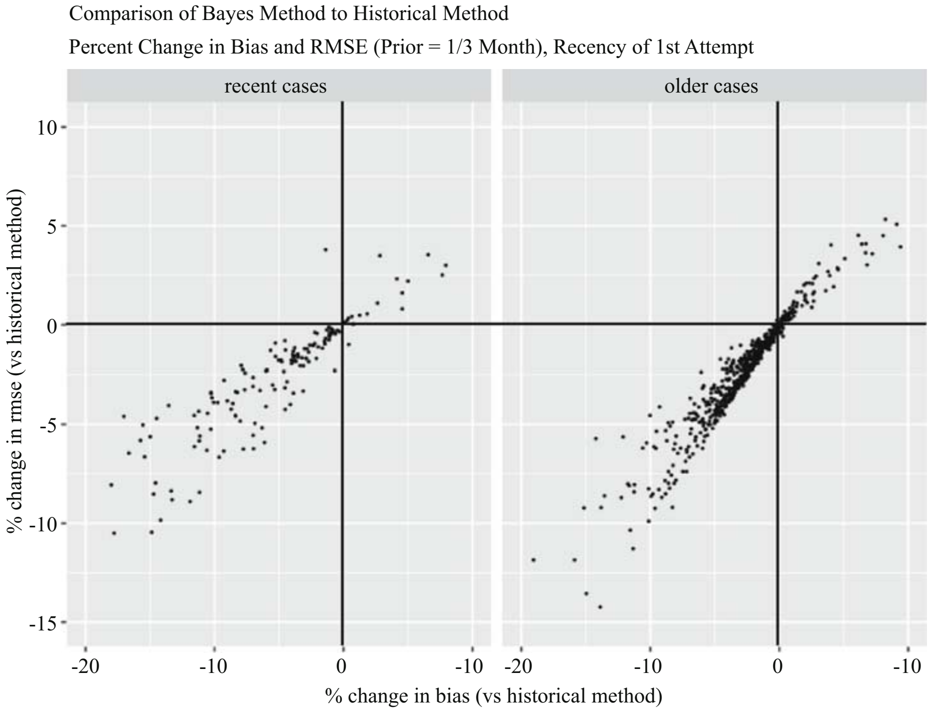


Fig. 6. Percent change in RMSE and bias of bayesian method versus. historical method by recency of first attempt.

6. Discussion and Future Work

This article discussed several methods for generating predictions for the estimated lag (in days) between first attempt and first contact for a case in the NHIS. The results showed that using model-based methods that leverage data external to the current data collection, in this case historical survey data, can be useful for improving predictions in the current month.

Because of the time-varying nature of the "day of the first attempt covariate", the *historical method* and the *Bayesian method* were able to capture that variability, resulting in better predictions, particularly near the point of first attempt. Further, the model-based methods improved estimates with relatively few covariates, including aggregate geographic information, basic interviewer information, and the day of first attempt. Improvements could be more significant with more information, either for individual cases in a longitudinal setting, or a richer auxiliary frame data in a cross-sectional setting.

The *Bayesian* method provided modest improvements over the *historical* method, with the largest gains visible early in the data collection period. Additionally, the *Bayesian* method provided small improvements in the identification of cases with excessive lags at a higher sensitivity than the other methods, particularly when weak priors were used. When sensitivity and specificity were considered equally in an application of the use of predictions, the *Bayesian* method performed slightly better early in data collection, though the *historical* method was competitive. In general, early in data collection the benefit of the Bayesian framework is that external data can be incorporated into the prediction process as priors. Those priors can help improve the stability of predictions when working with the partially accumulated data that exist early in the data collection period. Later in data collection, the likelihood (data from the in-progress period) takes over, reducing bias that might have been introduced by considering historical data only, although the prior continues to reduce the variance of the predictive coefficients, stabilizing the posterior predictions of expected lag. The *Bayesian* method also offers flexibility in how much influence the prior has on the posterior prediction, in the form of the constant, c.

There are limitations to the model-based predictions. From a data quality perspective, the NOI and the CHI data are self-reported by interviewers, who are expected to record outcomes of contact attempts immediately after they occur. West and Kreuter (2013), Biemer et al. (2013), and Bates et al. (2010) have discussed potential quality issues in self-reported interviewer paradata, due to under-reporting of undesirable behaviors, or the error inherent in guesswork. Additionally, as shown in the comparison of Figure 2 and Figure 3, predictions can be more inaccurate the farther in time away from the first attempt for a case, meaning the covariates currently in the model underestimate the length of the lag for cases with long lags. This may be for one of two reasons. First, in the hurdle portion of the model, estimated by logistic regression, the discrimination between sample units with zero lag and those with positive lags is not perfect. As a result, some units for which we predict a zero lag with high confidence may, in fact, have a long lag, and the high probability of a zero lag would cause our predictions to underestimate the expected lag. Then, in the Weibull portion of the model, which estimates lag length for non-zero lags, the estimate of the shape parameter may also underestimate the actual lag.

Future work could extend into several different areas. First, we used a standard parametric Weibull model to predict time (days) to contact. Models for integer-coarsened data, perhaps as a special case of interval-censored data (Sun 2006), could be considered, although we believe the impact of this level of refinement would be relatively minor. More work could be done to append auxiliary variables onto the survey data, and potentially include a flexible set of model covariates over time, in order to improve model fit and subsequent predictions. Household-level data may be more predictive than aggregate geographic information from the PDB. Additionally, this method should be replicated for different estimators in different data collection settings. Selected data collection parameters and associated models would likely be different across surveys with different data collection periods, modes, or data collection features. In situations where time-to-event modeling is useful or appropriate for estimating selected data collection parameters, it may be useful to explore additional tools for identifying the appropriate survival model, including non-parametric methods, such as Kaplan-Meier curves, in addition to the goodness-of-fit tests used in Subsection 3.1.

More benefit could be extracted from the Bayesian framework, as well. Here, we leveraged the Bayesian framework primarily to combine both historical and current accumulating data to maximize the information used in our predictions of posterior coefficient estimates. But after that, we only used the point estimate of the predicted lag for evaluating our predictions or identifying short and long lags in the online supplementary application. We could also leverage the credible interval around predictions for evaluation or application. In the evaluation phase, we could evaluate whether the predicted credible interval for a case covers its actual lag, as opposed to only looking at the distance between the predicted point estimate and actual lags. Similarly, we could rank cases that have minimally overlapping credible intervals by predicted lags and evaluate whether those rankings match the rankings in the actual lags. This would help us understand whether our predictions of lag were biased downward but ordered correctly. Additionally, we could evaluate the accuracy of predictions for cases with narrow credible intervals versus wide intervals, to better understand if there are domains or types of cases for which our predictions are more accurate than others.

In an application setting, the credible interval could be used to help identify cases for intervention. In the case where a threshold is used (e.g., lags of at least a certain number of days are considered long), the credible interval could be used to determine whether a case met that threshold. For example, in the case where a point estimate met the threshold, but a large portion of the credible interval did not meet the threshold, we may not consider that case to have a long lag. The additional information provided by the credible interval would help identify the cases likely to have long lags in a more complete way. One could also consider the distribution of rank order, for example focusing on cases that we find to be likely to have the longest times to contact for intervention if resources are limited for such intreventions. The strategy for assigning a value to c is also important. For this application, we varied the value of c, allowing the priors to represent a range of 1/3 of a month of data to three months of data for the results in order to demonstrate the effect of prior data strength on the posterior predictions of lag. However, the Bayesian framework could be extended by putting a prior on the value for c using power prior methodology (Ibrahim and Chen 2000; Hobbs et al. 2011). Power priors provide a statistical framework for weighting priors in a more sophisticated way than we did for this work.

Finally, it is important to consider how to utilize these improved predictions during data collection. Although this article has focused on the technical issue of predicting the day lag between first attempt and first contact, there is a broader issue of exactly how these results could be used to affect action in the field. There are at least two major categories of actions that could be taken. The first is load leveling. Given the extra effort that a successful contact will entail, one could ensure that interviewers are assigned a reasonable mix of cases that will allow them to get through a workday without leaving a large number of uncontacted cases because they had too many likely responders on a given day or end up short of cases to work because they had too many unlikely responders. The second category follows adaptive or responsive design interventions. For example, if time-to-contact was uncorrelated with response propensity or outcomes of interest, then at some point during data collection, cases could be dropped from operations if the estimated time-to-contact was too long. Alternatively, if time-to-contact was negatively correlated with response propensity, or correlated in any way with outcomes of interest, cases with longer predicted time-to-contact could be assigned high priority early in the collection period so that interviewers spent more time on those cases, or have alternative data collection strategies applied, such as assigning the case to multiple interviewers, leaving materials that might improve the chance for future contact, and so forth.

7. References

Bates, N., J. Dahlhamer, P. Phipps, A, Safir, and L. Tan. 2010. “Assessing Contact History Paradata Quality Across Several Federal Surveys,” In Proceedings of the American Statistical Association 2010 Joint Statistical Meeting, Vancouver, Canada. Available at: http://www.asasrms.org/Proceedings/v2010f.html (accessed July 2017).

Biemer, P., P. Chen, and K. Wang. 2013. “Using Level-Of-Effort Paradata in Non-Response Adjustments with Application to Field Surveys.” *Journal of the Royal Statistical Society* A176: 147–168. DOI: https://doi.org/10.1111/j.1467-985X.2012.01058.x.

Biffignandi, S., and J. Bethlehem. 2021. Web Surveys and Other Modes of Data Collection. In *Handbook of Web Surveys*. DOI: https://doi.org/10.1002/9781119371717.ch6.

Calinescu, M., S. Bhulai, and B. Schouten. 2013. “Optimal Resource Allocation in Survey Designs,” *European Journal of Operations Research* 226: 115–121. DOI: https://doi.org/10.1016/j.ejor.2012.10.046.

Census Bureau. 2008. *A Compass for Understanding and Using American Community Survey Data*. Available at: https://www.census.gov/content/dam/Census/library/publications/2008/acs/ACSGeneralHandbook.pdf (accessed October 2017).

Census Bureau. 2016. *Planning Database Documentation.* Available at: https://www.census.gov/research/data/planning_database/2016/docs/PDB_Block_Group_2016-07-28a.pdf (accessed July 2017).

Chesnut, J. 2013. “Model-based mode of data collection: switching from internet to mail in the American Communities Survey.” *2013 American Communities Survey Research and Evaluation Report Memo Series* #ACS-13RER, 18, 1–17. Available at: https://www.census.gov/library/working-papers/2013/acs/2013_Chesnut_01.html (accessed July 2017).

Christy, J. 2014. “Use of Response Propensity Scores to Direct CAPI Field Activity.” In Proceedings of the American Association for Public Opinion Research Annual Conference, May 15–18, Anaheim, CA, USA. DOI: https://doi.org/10.1111/j.1467-985X.2012.01058.x.

Coffey, S., B. Reist, and P.V. Miller. 2019. "Interventions On-Call: Dynamic Adaptive Design in the 2015 National Survey of College Graduates," *Journal of Survey Statistics and Methodology* 8: 726–747. DOI: https://doi.org/10.1093/jssam/smz026.

Couper, M.P. 2000. "Usability Evaluation of Computer-Assisted Survey Instruments," S*ocial Science Computer Review* 18: 384–396. DOI: https://doi.org/10.1177/089443930061800402.

Couper, M.P. 2017. "Birth and Diffusion of the Concept of Paradata " *Advances in Social Research* 18: 14–26. Available at: https://jasr.or.jp/english/JASR_Birth%20and%20-Diffusion%20of%20the%20Concept%20of%20Paradata.pdf (accessed October 2020).

Durrant, G.B., and F. Steele. 2009. "Multilevel Modelling of Refusal and Non-Contact in Household Surveys: Evidence from Six UK Government Surveys," *Journal of the Royal Statistical Society* A172: 361–381. DOI: https://doi.org/10.1111/j.1467-985X.2008.00565.x.

Edwards, B., H. Sun, and R. Hubbard. 2020. "Behavior change techniques for reducing interviewer contributions to total survey error." In *Interviewer Effects from a Total Survey Error Perspective*: 77–90. Chapman and Hall/CRC.

Feng, C.X. 2021. "A comparison of zero-inflated and hurdle models for modeling zero-inflated count data," *Journal of Statistical Distributions and Applications* 8: 8. DOI: https://doi.org/10.1186/s40488-021-00121-4.

Groves, R.M., and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys*, New York: Wiley.

Groves, R.M., and S.G. Heeringa. 2006. "Responsive Designing for Household Surveys: Tools for Actively Controlling Survey Errors and Costs," *Journal of the Royal Statistical Society* A169: 439–457. DOI: https://doi.org/10.1111/j.1467-985X.2006.00423.x.

Hobbs, B.P., B.P. Carlin, S.J. Mandrekar, and D.J. Sargent. 2011. "Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials." *Biometrics* 67: 1047–1056. DOI: https://doi.org/10.1111/j.1541-0420.2011.01564.x.

Ibrahim, J.G., and M.H. Chen. 2000. "Power prior distributions for regression models." *Statistical Science* 15: 46–60. DOI: https://doi.org/10.1214/ss/1009212673.

Jackson, M.T., C.B. McPhee, and P.J. Lavrakas. 2020. "Using Response Propensity Modeling to Allocate Noncontingent Incentives in an Address-Based Sample: Evidence from a National Experiment," *Journal of Survey Statistics and Methodology* 8: 385–411, DOI: https://doi.org/10.1093/jssam/smz007.

Laflamme, F., and M. Karaganis. 2010. "Implementation of Responsive Collection Design for CATI Surveys at Statistics Canada." In Proceedings of the Q2010, 3–6 May, Helsinki, Finland. Available at: https://www.researchgate.net/profile/Francois-Laflamme-3/publication/228583181_Implementation_of_Responsive_Collection_Design_for_CATI_Surveys_at_Statistics_Canada/links/5436b2120cf2dc341db46d70/Implementationof-Responsive-Collection-Design-for-CATI-Surveys-at-Statistics-Canada.pdf (accessed July 2017).

Luiten, A., J. Hox, and E. de Leeuw. 2020. "Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and

Surveys," *Journal of Official Statistics* 36(3): 469–487. DOI: https://doi.org/10.2478/-JOS-2020-0025.

Ma, L., M. Yan, and J. Weng. 2015. "Modeling Traffic Crash Rates of Road Segments through a Lognormal Hurdle Framework with Flexible Scale Parameter," *Journal of Advanced Transportation* 49: 928–940.

Mittereder, F., and B.T. West. 2021. "A Dynamic Survival Modeling Apporach to the Prediction of Web Survey Breakoff." *Journal of Survey Statistics and Methodology* 10: 945–978. DOI: https://doi.org/10.1093/jssam/smab015.

Mneimneh, Z., L. Lyberg, S. Sharma, M. Vyas, D.B. Sathe, F. Malter, and Y. Altwaijri. 2018. "Case Studies on Monitoring Interviewer Behavior in International and Multinational Surveys." In *Advances in Comparative Survey Methods*. John Wiley & Sons. DOI: https://doi.org/10.1002/9781118884997.ch35.

Mullahy, J. 1986. "Specification and Testing of some Modified Count Data Models," *Journal of Econometrics* 33: 341–365. DOI: https://doi.org/10.1016/0304-4076(86)90002-3.

NCHS, National Center for Health Statistics. 2018. *National Health Interview Survey: Survey Description*. Technical Report. Available at: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2017/srvydesc.pdf (accessed July 2017).

Peytchev, A., J. Rosen, S. Riley, J. Murphy, and M. Lindblad. 2010. "Reduction of Nonresponse Bias through Case Prioritization." *Survey Research Methods* 4: 21–29. DOI: https://doi.org/10.18148/srm/2010.v4i1.3037.

Peytchev A. 2014. *Models and Interventions in adaptive and responsive survey designs*. DC-AAPOR Panel on Adaptive Survey Design. Washington D.C. Available at: http://dc-aapor.org/ModelsInterventionsPeytchev.pdf (accessed July 2017).

Peytchev, A., D. Pratt, and M. Duprey. 2020. "Responsive and Adaptive Survey Design: Use of Bias Propensity During Data Collection to Reduce Nonresponse Bias," *Journal of Survey Statistics and Methodology* 10: 131–148. DOI: https://doi.org/10.1093/jssam/smaa013.

Roberts, C., C. Vandenplas, and M.E. Stahli. 2014. "Evaluating the Impact of Response Enhancement Methods on the Risk of Nonresponse Bias and Survey Costs." *Survey Research Methods* 8: 67–80. DOI: https://doi.org/10.18148/srm/2014.v8i2.5459.

Rose, C., S. Martin, K. Wannemuehler, and B. Plikaytis. 2006. "On the Use of Zero-Inflated and Hurdle Models for Modeling Vaccine Adverse Event Count Data," *Journal of Biopharmaceutical Statistics* 16: 463–481. DOI: https://doi.org/10.1080/10543400600719384.

Schouten, B., F. Cobben, and J. Bethlehem. 2009. "Indicators for Representativeness of Survey Response." *Survey Methodology* 39: 29–58. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2009001/article/10887-eng.pdf?st=Ox4bRwqM.

Schouten, B., N. Shlomo, and C. Skinner. 2011. "Indicators for Monitoring and Improving Representativeness of Survey Response," *Journal of Official Statistics* 27(2): 231–253. Available at: https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/indicators-for-monitoring-and-improving-representativeness-of-response.pdf.

Schouten, B., N. Mushkudiani, N. Shlomo, G. Durrant, P. Lundquist, and J. Wagner. 2018. "A Bayesian Analysis of Design Parameters in Survey Data Collection," *Journal of*

Survey Statistics and Methodology 6: 431–464: DOI: https://doi.org/10.1093/jssam/smy012.

Smithson, M., and E. Merkle. 2013. “Chapter 5: Count Variables,” In *Generalized Linear Models for Categorical and Continuous Limited Dependent Variables*. New York: Chapman and Hall/CRC. Print.

Sun, J. 2006. *The Statistical Analysis of Interval-Censored Failure Time Data* New York: Springer.

Tolliver, K., J. Fields, S. Coffey, and A. Nagle. 2019. “Combating Attrition Bias Using Case Prioritization in the Survey of Income and Program Participation.” In Proceedings from the 2019 AAPOR Conference, Toronto, Ontario, USA. Available at: http://www.asasrms.org/Proceedings/y2019/files/1199523.pdf (accessed July 2017).

Wagner, J., and T.E. Raghunathan. 2010. “A New Stopping Rule for Surveys.” *Statistics in Medicine* 29: 1014–1024. DOI: https://doi.org/10.1002/sim.3834.

Wagner, J., and F. Hubbard. 2014. “Producing Unbiased Estimates of Propensity Models during Data Collection,” *Journal of Survey Statistics and Methodology* 2: 323–342. DOI: https://doi.org/10.1093/jssam/smu009.

Wagner, J., B.T. West, N. Kirgis, J.M. Lepkowski, W.G. Axinn, and S.K. Ndiaye. 2012. “Use of Paradata in a Responsive Design Framework to Manage a Field Data Collection.” *Journal of Official Statistics* 28: 477–499. Available at: https://www.scb.se/contentassets/f6bcee6f397c4fd68db6452fc9643e68/use-of-paradata-in-a-responsive-design-framework-to-manage-a-field-data-collection.pdf (accessed July 2017).

Walejko, G., and J. Wagner. 2018. “A Study of Interviewer Compliance in 2013 and 2014 Census Test Adaptive Designs” *Journal of Official Statistics* 34(3): 649–670. DOI: https://doi.org/10.2478/jos-2018-0031.

Walsh, R., and J. Coombs. 2013. “Case Reassignment: When making contact is a two-person job.” In Proceedings of the Joint Statistical Meetings, Boston, MA, USA: 3690–3700. Available at: https://www.researchgate.net/publication/296847275_Case_Reassignment_When_making_contact_is_a_two-person_job (accessed July 2017).

West, B.T., and R.M. Groves. 2013. “The PAIP Score: A Propensity-Adjusted Interviewer Performance Indicator,” *Public Opinion Quarterly* 77: 352–374. DOI: https://doi.org/10.1093/poq/nft002.

West, B.T., and F. Kreuter. 2013. “Factors Affecting the Accuracy of Interviewer Observations: Evidence from the National Survey of Family Growth (NSFG).” *Public Opinion Quarterly* 77: 522–548. DOI: https://doi.org/10.1093/poq/nft016.

West, B.T., J. Wagner, S. Coffey, and M.R. Elliott. 2021. “Deriving Priors for Bayesian Prediction of Daily Response Propensity in Responsive Survey Design: Historical Data Analysis vs. Literature Review.” *Journal of Survey Statistics and Methodology* 11: 367–392. DOI: https://doi.org/10.1093/jssam/smab036.

Williams, D., and J.M. Brick. 2018. “Trends in U.S Face-to-Face Household Survey Nonresponse and Level of Effort.” *Journal of Survey Statistics and Methodology* 6: 186–211. DOI: https://doi.org/10.1093/jssam/smx019.

Received April 2022
Revised October 2022
Accepted January 2023

Journal of Official Statistics, Vol. 39, No. 3, 2023, pp. 351–379, http://dx.doi.org/10.2478/JOS-2023-0016

Towards Demand-Driven On-The-Fly Statistics

Tjalling Gelsema[1] *and Guido van den Heuvel*[1]

A prototype of a question answering (QA) system, called Farseer, for the real-time calculation and dissemination of aggregate statistics is introduced. Using techniques from natural language processing (NLP), machine learning (ML), artificial intelligence (AI) and formal semantics, this framework is capable of correctly interpreting a written request for (aggregate) statistics and subsequently generating appropriate results. It is shown that the framework operates in a way that is independent of a specific statistical domain under consideration, by capturing domain specific information in a knowledge graph that is input to the framework. However, it is also shown that the prototype still has its limitations, lacking statistical disclosure control. Also, searching the knowledge graph is still time-consuming.

Key words: Dissemination; artificial intelligence; question answering; text-to-SQL; information modeling.

1. Introduction

The usual way a national statistical institute (NSI) disseminates most of its output to the general public is through an output database, generally made available through the NSI's website. Within Statistics Netherlands (SN) for instance, StatLine is the output facility in which most of the statistics compiled are published. On SN's website, StatLine is accompanied by searching and selection facilities in order to aid the general public in finding statistics of interest. Other NSIs have similar facilities.

These output databases have in common that they store pre-compiled statistics: numbers that are calculated in advance of a public's request for information. This type of generating output can therefore be called supply-driven: the assumption is that the vast amount of statistical tables that are compiled and disseminated is sufficient to serve the main public's interest.

In contrast to the supply-driven approach to output dissemination is the demand-driven approach. We characterize it here as the calculation of a statistic as the result of a user's request. Usually, output of this type is the result of customized research originating from a request for statistics not found on the website. While the usual supply-driven type of dissemination is fast in terms of the time it takes between a request and delivering output, the usual demand-driven type is slow due to the customized process of producing statistics.

To our knowledge, no NSI is currently capable of combining the advantages of demand-driven output generation with the fast access output databases offer. One notable exception

[1] Statistics Netherlands, Research and Development, Henri Faasdreef 312 The Hague 2492 JP, the Netherlands. Emails: te.gelsema@cbs.nl and g.vandenheuvel@cbs.nl

Acknowledgments: The authors are grateful to Dick Windmeijer and Joeri van Etten for their careful reading of an earlier version of this article, and for their useful comments. We very much appreciated the thorough and critical review of Peter-Paul de Wolf and implemented many of his valuable suggestions for improving the article. We thank the numerous anonymous referees for their much appreciated comments.

is the Australian Bureau of Statistics (ABS) with their TableBuilder initiative. This on-line tool allows users to create tables, graphs and maps of census data to a high level of detail.

However, more specifically, we think that the process of automatically interpreting a user's written request for statistics followed by the calculation of relevant output, for any domain, is currently implemented nowhere. In this article we introduce a framework for doing just that: demand-driven on-the-fly statistics.

We postulate that demand-driven on-the-fly statistics could offer a more personalized kind of statistics than the current types of dissemination do. Questions like "How many people live in my street?" and "How many people work in Leiden and live in The Hague?" can be relevant from a personal perspective and may very well be the kind users want instant response to. Currently they are not addressed by SN's output facility, due to (1) the level of detail requested (SN does not publish population figures at street level) and (2) the combination of sources required (SN does not publish figures combining population statistics with employer residence). SN however does posses the sources for calculating these figures and, assuming proper disclosure control, there is no reason not to publish them. Of course, with high levels of detail and with many combinations of data to account for, disclosure control of output data is a challenge.

To emphasize the potential innovative character of the framework we are proposing (in view of the current techniques for disseminating statistics) we list its initial goals: the framework should

1. translate a request in layman terms to the more technical statistical vocabulary,
2. understand a request to the extent that a query for calculating a result can be generated,
3. automatically determine which tables containing microdata should be combined to calculate the result,
4. calculate the result using microdata, and,
5. explain the result to the user.

In our view, the combination of the requirements above is unique and could be a driver for a more user-centered way of disseminating statistics.

In this article we describe the technicalities and the applications of a framework for the real-time calculation of statistical figures, in response to a written request. While on-the-fly disclosure control is a critical component for such a framework to be put into actual use, we exclude it in this article. First, there are numerous references to research on the topic (see, e.g., Thompson and Broadfoot et al. 2013; Marley and Leaver 2011; Fraser and Wooton 2006; Enderle and Giessing et al. 2006) and we have little to contribute on the matter, other than point the reader in the direction of the promising notion of differential privacy Ji et al. (2014). Second, even without on-the-fly disclosure control, the technicalities involved in our framework are nontrivial and, we feel, constitute a field of research that is relevant on its own.

At this point we must mention the development of the so-called Information Dialogue (ID) SN was pursuing: a voice assistant application for smart phone users that is capable of answering frequently asked questions on some topics of interest. As of January 2020, these topics included population, income and job market, the data for which is provided by SN's open data framework.

SN's ID differs from our approach in some notable ways. First, the ID focusses on sensible dialogues to direct a user to the information required, which takes a large burden on the development of the ID. Our approach is to take a best guess from a single question posed in natural language and subseqently generate an answer in one stroke instead of through a dialogue. It is therefore an example of a question answering (QA) system Indurkhya and Damerau (2010) rather than a dialogue system. Second, SN's open data framework provides exactly the same output tables as StatLine, so the ID can be seen as a convenient way to navigate through some of SN's output tables. It therefore does not follow the demand-driven approach that we discussed earlier. Third, it turns out that each topic requires a separate dialogue – one of the reasons dialogue design takes a large portion of the ID's development. Instead, our approach, we claim, is more generic: our framework is essentially agnostic of the statistical domain of a request; its mechanisms will thus interpret a question about the Dutch population in much the same way as a question about income. Fourth, since our approach is based upon aggregating and combining microdata to form an answer, we claim that the number of questions that can potentially be answered is much higher than if we used aggregated data as a starting point. We therefore claim that the questions our framework is capable of answering are potentially more numerous and more diverse than those of the ID (adding to the complexity of a potential solution for disclosure control).

To give an outline of the objectives of our framework called Farseer, consider Figure 1 below. At the lowest level of detail, it consists of four stages, two of which, that is, *Statistical disclosure control* and *Inform*, we will only briefly discuss in this article. The input of the first stage, that is, *Interpret*, is a request for statistics in written natural language form. Its output is a formal representation of this request, in the sense that this representation essentially contains all metadata (such as variables, selection criteria, aggregation instructions) needed to compute a figure, or a table of figures, as well as to explain the figure's meaning. The steps needed to perform *Interpret* include several natural language processing (NLP) tasks in part supported by machine learning (ML) routines. The details of *Interpret* are described in Subsections 2.2 to 4.4. The second stage, *Generate and execute query*, is a translation of the formal representation to an executable language; as of March 2020, both MySQL and Transact-SQL are supported. The details of *Generate and execute query* are explained in Subsection 2.5. The purpose of *Inform* is to translate the formal representation of *Interpret* back to a human readable form, informing the user of the exact meaning of the figure(s) computed. It is important to report on interpretation

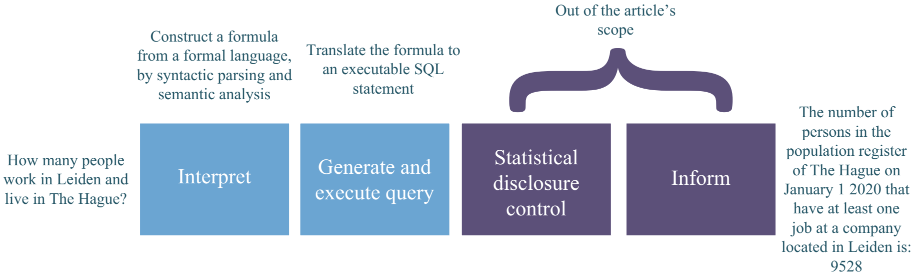

Out of the article's scope
Construct a formula from a formal language, by syntactic parsing and semantic analysis
Translate the formula to an executable SQL statement
How many people work in Leiden and live in The Hague?
Interpret
Generate and execute query
Statistical disclosure control
Inform
The number of persons in the population register of The Hague on January 1 2020 that have at least one job at a company located in Leiden is: 9528

Fig. 1. Overview of Farseer

issues due, for example, to possible translation differences between casual and formal speech (such as the translation between "live in The Hague" and "the population register of The Hague" in Figure 1) or of assumptions made (such as "January 1 2020" in Figure 1). At the moment, we are studying on developing algorithms for the *Inform* stage.

Farseer is an acronym of for Automated Replies to Statistical Estimation Enquiries and Requests. Full code is available at. github.com/tgelsema/farseer.

Scientifically, Farseer is an example of a text-to-SQL system (also known as NL2SQL); see Katsogiannis-Meimarakis and Koutrika (2021) for an overview of other such systems. Farseer uses a traditional, rule based approach for the actual translation, combined with ML for specific subtasks. In Section 4 we compare our approach with that of others.

The rest of this document is organized as follows. In Section 2 we present the technical details of the Farseer framework. Section 3 presents the results of an experiment in which the framework was applied to a statistical domain (viz. crime) that was new to the framework, in the sense that it differed from the domain used during development. These results are discussed in Section 4 and we conclude with some remarks on future research in Section 5.

2. Interpreting a Request for Statistics and Calculating its Result

In the following subsections, we describe in detail the mechanisms of the *Interpret* and *Generate and execute query* stages, which turn a request for statistics written in natural language into a formula first, before translating it to an executable query.

First, we must emphasize that Farseer currently is capable of interpreting Dutch requests only. We expect however that the mechanisms of interpreting an English request for instance can be developed in much the same way. Nevertheless, all the example requests presented in this section are in Dutch; we will give proper translations in English as we proceed.

The parts of the *Interpret* component are roughly those commonly found in NLP Indurkhya and Damerau (2010), though our interpretation of some of these parts is non-standard. In order to process a request it must be tokenized first, breaking up the request into individual words or composites of words. Consider Figure 2 where the English translation of the example request is "How many people live in The Hague?". The result of tokenization is a list of tokens; note that in this case it is recognized that "Den Haag" forms a single token instead of two.

The purpose of named entity recognition (NER) is to map certain tokens to prior knowledge. In our case this means that some tokens are related to vertices and edges in a knowledge graph, the details of which are described in Subsection 2.1. Note that in Figure 2 the token "den haag" is recognized as an object (an edge in this case) named "'s-Gravenhage"; it is the official name of the city of The Hague. Tokenization and NER are described in some detail in Subsection 2.2. The purpose of the lookup table mentioned in Figure 2 is to translate common language (such as "Den Haag") to official language ("'s-Gravenhage").

We found a need for performing some non-standard part-of-speech tagging (POS), that is, in order to classify requests into one of currently eleven categories. The classification process itself uses ML: given the output of POS applied to a request, our ML routine

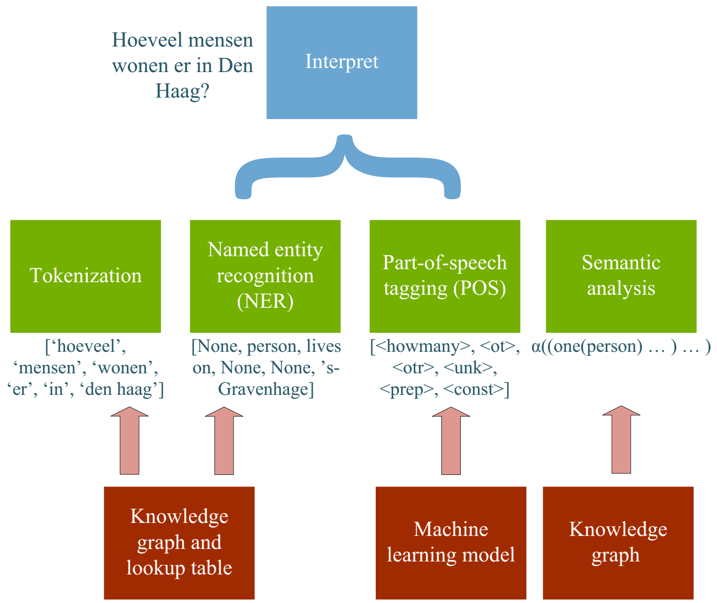

Hoeveel mensen wonen er in Den Haag?
Interpret
Tokenization
Named entity recognition (NER)
Part-of-speech tagging (POS)
Semantic analysis
['hoeveel', 'mensen', 'wonen', 'er', 'in', 'den haag']
[None, person, lives on, None, None, 's-Gravenhage]
[<howmany>, <ot>, <otr>, <unk>, <prep>, <const>]
α((one(person) …) …)
Knowledge graph and lookup table
Machine learning model
Knowledge graph

Fig. 2. Outline of the interpret component.

estimates the category corresponding to the request. These categories in turn give direction to the semantic analysis part of the *Interpret* component. The output of POS is a list of keywords (such as <*howmany*>, <*prep*> for prepositions, and <*unk*> for unknowns) of the same length as the list of tokens. This list can be seen as a structural excerpt of the request: all domain-specific information is intentionally left out. In this way the classification process becomes domain independent. POS and our classification routines are explained in Subsection 2.3.

In Subsection 2.4 semantic analysis is explained. Also this stage is developed in a non-standard way, in that it introduces a formal language customized for the capturing of statistical information. The output of semantic analysis is a formula of the language, which represents the formal translation of a request.

We conclude in Subsection 2.5 where we describe the translation of a formula to an executable query.

At this point we want to stress that our main contributions to the field of NLP applied to statistics, we feel, are the semantic analysis and the query generation stages. For this reason Subsections 2.1, 2.5 and in particular 2.4 contain more detail than others.

2.1. Knowledge Graphs

In this section we shall refer to a knowledge graph as a directed graph of which the labeled vertices and edges contain structural knowledge about a domain of statistical interest. We present here our own version of a knowledge graph, specifically designed for capturing statistical information. Other versions use, for example, linked open data or other approaches based on the Resource Description Framework (RDF) (see Hayes and Patel-Schneider 2014; Brickley and Guha 2014; Motik et al. 2012) but we found our particular version more expressive in a statistical context and better suited for our Farseer application. How our graphs should be interpreted is explained in more detail in Gelsema (2012): there the relationship between a knowledge graph and actual data is made precise by means of a semantic function that has as a domain (expressions over) a graph and as a codomain a collection of sets and functions that represent data(sets). We will return to the exact semantics of knowledge graphs in Subsection 2.4 and point out the difference with RDF-based graphs.

An example of a knowledge graph is given in Figure 3. Essentially, this graph (as all knowledge graphs of the kind we consider) stores five types of information; each type is given a different color in Figure 3.

First, the red vertices of the graph are (called) object types: these are types of statistical interest, such as *person* and *job* and they represent countable sets of objects of interest. Relationships between object types are the edges colored green: these are called object type relations and, for example, record the relationship between a job and the person that occupies it, or the relationship between a person and the address he or she lives on. Note that object type relationships have a direction: this indicates that, for example, associated with a job is exactly one person (and not necessarily the other way around). The blue-colored edges are variables: they go from an object type to a type which we call a phenomenon, for lack of a better term. Thus, the variable labeled *job title* assigns to each job a member of the type *occupations*: this is the functional perspective of a statistical variable (and an object type relation) proposed by Gelsema (2012) that we will follow here. Note the special variables labeled *one* and *all*; we assume we have a variable of both kinds available for each object type (but only some of them are shown in Figure 3). The variable *one* represents a function that assigns to each object of the corresponding type the number 1; we use this for counting objects later on. The variable *all* represents a function of which all objects of the corresponding type are mapped to the same object of the type 1; the latter type can be thought of as a singleton set {*} containing one fixed but unnamed member *. We use the *all* variable for grouping objects later on, see a “group by” statement from SQL. Some might prefer more ‘formal’ labels like *has adress* instead of

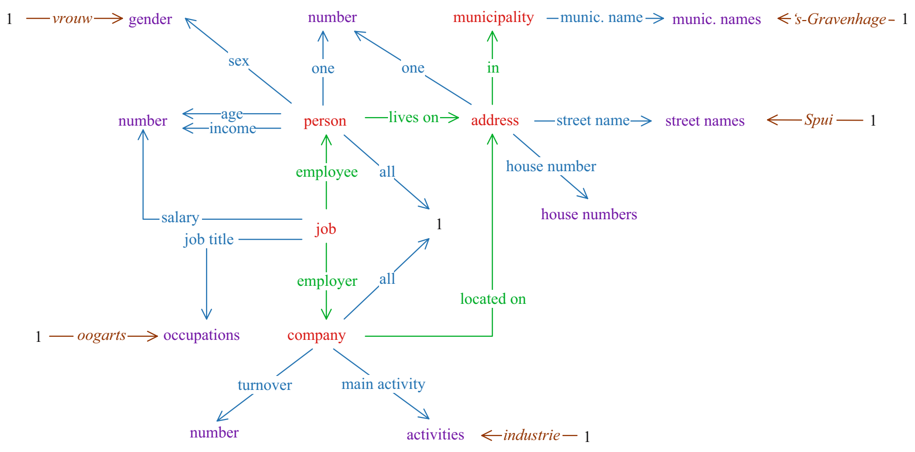


Fig. 3. A knowledge graph.

lives on. We stress however that the labels play no role in the NLP-part of Farseer though. Also, this example graph does not reflect the real world in all its details. For instance, a company might have more than one address. We stress that these details can be incoporated in the graph though.

Note that object type relations and variables, as in fact all edges in the graph of Figure 3, are subject to composition. By this we mean that, for example, the variable *municipality name of the place of residence of a person* is a variable that is composed of the object type relationships *lives on* and *in* and the variable "municipality name" by following the edges in the given direction Gelsema (2012) so that the sets-and-functions interpretation of composing edges corresponds with the usual composition of functions.

The brown edges are called constants: they originate from the special type denoted 1 and have a phenomenon type as a target. This is a convenient way to indicate that, for example, *vrouw* (which translates to *female*) is a member of the type *gender*, because functions v of the form $v : \{*\} \to a$ are in one-to-one correspondence with members of the set a. Note that we did not translate the labels of the constants in Figure 3; this is because they are relevant in the interpretation process, which, as mentioned, has a Dutch request as input. The label *industrie* translates to *industry*, *oogarts* translates to *ophthalmologist* (or just *eye doctor*) and *'s-Gravenhage* and *Spui* translate to themselves. Note also that the term phenomenon might give the reader other connotations than intended: a phenomenon must be seen as a set of constants, as explained above, nothing more and nothing less. In the example knowledge graphs in this paper, the constants are invariably categories from a classification, so a synonym for phenomenon might be just classification or code list. We sometimes use *type* to refer to a vertex (label) and element to refer to an edge (label), see Gelsema (2019).

The graph in Figure 3 (or more precise: its Dutch translation) was used in various versions during the development of Farseer. We note that Figure 3 depicts a summary: the actual graph we used has many more constants. Also, in Figure 3 for convenience the types **1** and *number* have more than one occurrence; in an actual graph there is just one occurrence of both. Routines for creating graphs of the kind of Figure 3 have been assembled in a Python module of the Farseer framework. To create an actual graph, its (colored) vertices and edges must be constructed by hand. Of course, the idea is that to accommodate a new statistical domain, new vertices and edges can be added to an already existing knowledge graph, thereby promoting reuse.

2.2. Tokenization and Named Entity Recognition

Tokenization is the process of breaking up a sentence into tokens: individual words or composites of words. During NER certain tokens are mapped to objects of interest; in our case these objects are vertices and edges from the knowledge graph explained in the previous section.

The basic procedure of tokenization, at least the one outlined in this article, is straightforward: treat individual words as tokens, ignore punctuation marks such as commas, periods, question marks and exclamation marks and bring every word to lowercase. This obviously goes wrong when a group of words such as "Den Haag" is encountered in a request. Therefore such composites of words are recorded in a lookup table: this table maps these to objects (vertices and edges) in the knowledge graph. In fact, since this is an instance of NER, the lookup table serves as the mapping mechanism of single words to objects as well. Tokenization and NER are thus performed in one stroke and the output of this combined procedure is a list of tokens and a list of objects of the same length.

Consider Figure 4 below where the result of tokenization and NER of our example request is given. The lookup table shows three entries: the token "mensen" (people) is mapped to the object type *person*, the token "wonen" is mapped to the object type relation *lives on* and the token "den haag" is mapped to the constant *'s-Gravenhage*. The rest of the tokens have no entry in the lookup table, so the result of NER in these cases is the (Python) keyword "None". Note that tokens can be mapped to vertices as well as edges. Also note that the lookup table serves as a translation mechanism between casual speech (den haag) and formal speech ('s-Gravenhage).

An actual lookup table can be large: if all vertices and edges in a knowledge graph need to be referenced, its size equals the size of the graph roughly. The lookup table we used during the development of Farseer was constructed by hand, aided partially by routines to account for the many constants.

Maintaining a lookup table can be labor intensive, especially if all synonyms of a word must be included as a key in the table as well. To keep the number of entries in the lookup table relatively low, we first turn to look for synonyms before consulting the lookup table. We roughly estimate that this reduces the lookup table to a third of its size. Further reduction can be accomplished if, during tokenization, tokens are converted to their root form – a process called stemming or lemmatization – so that there is no need to include conjugates and plurals in the lookup table.

Word embeddings Mikolov et al. (2013) are vector representations of words in a vocabulary that are computed from a corpus of text in such a way that words that can be exchanged for one another within a context of other words have embeddings that are near to each other given a certain measure of distance. Hence, words that are synonyms have embeddings at close distance, and this also holds for plurals and singulars, and conjugates of a verb. Unfortunately, the same is true for antonyms, so that word embeddings must be used with care. To keep the number of entries in the lookup table as low as possible, we used the following procedure. The word embedding is taken for each token (except for the most commonly occurring stop words) in the list of tokens that is not mapped directly to an

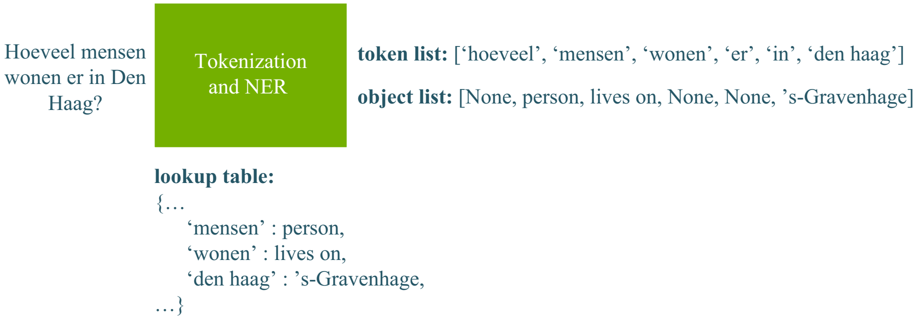

Hoeveel mensen wonen er in Den Haag?
Tokenization and NER
token list: ['hoeveel', 'mensen', 'wonen', 'er', 'in', 'den haag']
object list: [None, person, lives on, None, None, 's-Gravenhage]
lookup table:
{…
'mensen' : person,
'wonen' : lives on,
'den haag' : 's-Gravenhage,
…}

Fig. 4. Simple tokenization and NER.

object through the lookup table. Then the distance, that is, the cosine similarity, between this embedding and all embeddings of entries in the lookup table is computed. If this distance is above a certain threshold, then the corresponding entry in the lookup table is taken as a synonym. In this way, (pseudo) synonyms like ‘onderneming’ (company) for ‘bedrijf’ (business) are detected and handled properly. The threshold we took was 0.6. By experimenting, this was the optimum in that it detected proper synonyms as much as possible and avoiding improper synonyms.

The word embeddings we used were taken from the full corpus of Dutch wikipedia pages and precomputed by Bojanowski et al. (2017) using fastText. As a vocabulary, we took the 150,000 most occurring words from the text corpus. The dimensionality of the vectors used was 300. We omit the details of computing word embeddings and distance measures; the reader is referred to Mikolov et al. (2013) for a broad discussion.

If a token does not match a word in the lookup table, and it does not match a word in the full vocabulary, then it is substituted by a word in the lookup table with a minimal (normalized) Damerau-Levenshtein distance, upto a certain threshold. This corrects for small typographic mistakes in the request.

2.3. Part-Of-Speech Tagging and Request Classification

To aid the process of semantic analysis, we found it useful to classify a request based on the type of query that it requires, that is, whether it is a request for a total, an average, and so on. A template for the formula that is generated during semantic analysis can then be derived from the type of query.

Farseer was initially developed using rule-based methods to classify a request into one of currently eleven classes. It was found though that perfecting and maintaining these rules was cumbersome, so new ways were explored as classification mechanisms. We found that ML methods constitute a reasonable alternative.

The eleven classes we distinguish are exemplified in Table 1. They reflect the types of requests expected to be most common. They are to be interpreted as follows. Class 1 requests are requests for variables defined on individuals: persons, companies, etcetera. Class 2 requests require counting of individuals, as do class 7 requests which require calculating maximum counts. Class 3 requests require calculating a total of some numeric variable, as do class 8 requests, which maximize these totals. Class 4 requests require taking the average of some numerical variable; class 9 requests requires maximizing this average. Class 5 requests require calculating an average as a result of dividing two counts and class 11 requests require maximizing over this average. Class 6 requests require an interpretation of maximum (minimum); in the example request of Table 1 ‘biggest company’ could mean largest company according to turnover or the number of employees for instance. Class 10 requests require maximizing over a numerical variable.

Hence, classes 6 – 11 require taking a maximum (or minimum) of some value, while classes 1 – 5 do not. Also, classes 1, 6 and 10 do not require aggregation, so in an actual implementation they must be switched on only with very strict measures for disclosure control, or just switched off for privacy considerations. Also, calculating maxima or minima is generally considered to be unsafe, since these can reveal information about individuals.

Table 1. Request classes.

Class	Example request	English translation	Description
1	De adressen van bedrijven waar griffiers werken	The addresses of companies where registrars work	Properties of individuals
2	Hoeveel banen zijn er in Rotterdam?	How many jobs are there in Rotterdam?	Counts of individuals
3	Totale omzet in de bouwnijverheid	Total turnover for the construction industry	Totals for a single variable
4	Gemiddelde leeftijd van personen per gemeente	Average age of persons per municipality	Averages for a single variable
5	Gemiddeld aantal personen op een adres in Den Haag	Average number of people per address in The Hague	Average number of individuals
6	Wat is het grootste bedrijf in Delft?	What is the biggest company in Delft?	The greatest (or smallest) individual, according to the maximum of a preset variable
7	Welke gemeente heeft in totaal de meeste inwoners?	Which municipality has the most inhabitants in total?	The greatest (or smallest) individual, according to a maximum count
8	In welke gemeente in de totale omzet in de industrie het grootst?	Which municipality has the highest total turnover in industry?	The greatest (or smallest) individual, according to a total for a variable
9	Gemeente met het hoogste gemiddelde inkomen van vrouwen	Municipality with the highest average income earned by females	The greatest (or smallest) individual, according to a maximum average for a variable
10	Gemeente met de oudste inwoner	Municipality with the oldest inhabitant	The greatest (or smallest) individual, according to the maximum of a variable
11	Gemeente met gemiddeld het grootste aantal banen per bedrijf	Municipality with, on average, the highest number of jobs per company	The greatest (or smallest)individual, according to a maximum average number of other individuals

Note that requests could yield a single piece of information (one number or otherwise) or a table. For instance, the example request for class 3 in Table 1 should yield a single total, while the example request for class 4 asks for an average for each municipality, which then forms the dimension of a resulting table of averages. It should be stressed though that each class that requires aggregation includes requests that yield a dimensional table as well as a single total or average. Also note that some requests require selections, such as the example request for class 3, while others do not, such as the example request for class 4. All classes though include requests with one or more selection conditions. Here, companies are selected that have construction as their main activity.

The kind of POS tagging we propose in this article is non-standard and more straightforward than standard approaches Indurkhya and Damerau (2010), but nevertheless effective for our purpose of classifying a request into one of the eleven classes. Input to our POS tagging routine are the lists of tokens and objects mentioned in Subsection 2.2; the output is a list of keywords of the same length. The keywords we consider here are listed in Table 2. A keyword at some position in the list of keywords is derived from a token in the same position in the list of tokens, or from an object (vertex or edge) in the same position in the list of objects.

Thus, for the example request "Hoeveel mensen wonen in Den Haag" which is associated through tokenization with the list of tokens ['hoeveel', 'mensen', 'wonen', 'in' 'den haag'] and through NER with the list of objects [None, person, lives on, None, 's-Gravenhage], the list of keywords [<*howmany*>, <*ot*>, <*otr*>, <*prep*>, <*const*>] is derived. Note that the (single) token *wonen* is associated with the (single) object type relation *lives on* of which the name consists of two words. Note also that for each object (other than 'None') the corresponding keyword refers to its 'color' (<*ot*>, <*otr*> or <*const*>); the rest of the keywords is derived from their corresponding tokens.

A list of keywords associated with a request forms the input to our classification mechanism; its output is the estimated class for the request, according to Table 1. We

Table 2. Keywords in POS tagging.

Keyword	Token or object	English translation
< *with* >	met	with
< *prep* >	bij, op, in, van, uit	other prepositions
< *num* >	aantal, aantallen	number, numbers
< *howmany* >	hoeveel, vaak	how many, often
< *tot* >	totale, totaal	total
< *avg* >	gemiddelde, gemiddeld	average
< *all* >	ieder, iedere, elk, elke, al, alle	every, all
< *whowhat* >	welk, welke, wat, wie, waar	which, what, who, where
< *per* >	per, naar, voor, over	per
< *greatest* >	grootst, grootste, hoogst, hoogste, maximum, maximaal, maximale	greatest, highest, maximum, maximal
< *smallest* >	kleinst, kleinste, laagst, laagste, minimum, minimaal, minimale	smallest, lowest, minimum, minmal
< *most* >	meest, meeste	most
< *least* >	minst, minste	least
< *ot* >	**object type**	
< *otr* >	**object type relation**	
< *const* >	**constant**	
< *numvar* >	**numerical variable**	
< *catvar* >	**non-numerical variable**	
< *unk* >	all other tokens	

briefly discuss the application of a ML algorithm for our classification step. We refer to Chollet (2018) for the technicalities involved.

Our ML application has a sequential model of four layers, of which a one dimensional convolution (Conv1D) layer is the most important: such a layer is suitable for interpreting one dimensional sequential data, as are our keyword lists. As parameters for the Conv1D layer we took a window size of seven (data is read seven sequential keywords at a time) and we gave the layer 32 output units. Inputs are represented as a 20-dimensional vector of numerical labels corresponding to the keywords of Table 2, so requests are cut off at the twentieth keyword. As input layer an embedding is chosen with 32 dimensions. The output of the model is a number for the class that is estimated, along the lines of Table 1. The layers after the Conv1D are a max pooling layer followed by a fully connected layer with a softmax activation. See Géron (2017) for such an architecture.

This model was trained in supervised mode on some 2,155 handwritten and hand-labeled requests (which were subsequently transformed into keyword lists), of which 20% was used for validation. After 80 epochs of training with batch size 20, the accuracy of the model turned out to be about 96%. This number is unusually high and should deserve an explanation.

First, the requests we took as input came from a single source. This means that the model is biased towards his or her personal style of formulating a request. We expect that a mix of sources will decrease the accuracy of the model, but we expect it to perform better in environments other than our experimental. On the other hand, the 2,155 requests were intentionally subdivided into so-called chapters; each chapter consisted of 10–15 requests that were mere reformulations of one another, and each request in a chapter should therefore have identical semantics. In other words, care was taken to obtain a mix of alternatively formulated requests. Second, the classification of requests depends heavily on occurrences of words like ‘average’, ‘total’, or ‘greatest’, so that the model effectively learns to classify along these occurrences. In other words, the classification problem might not be that involved after all.

2.4. Semantic Analysis

The goal of semantic analysis is to assign meaning to a sentence through a formula or a term from some formal language. Usually, a first-order language Andrews (2002) is taken; other approaches use lambda calculus Barendregt (1984) as a formalism.

The language we used for semantic analysis was specifically designed to capture the meaning of statistical requests and is based on neither; at most it resembles relational algebra Codd (1970). It originated in Gelsema (2008) and took its algebraic form in Gelsema (2012) where the properties of an aggregation operator were investigated. In Gelsema (2019) the language was augmented with a selection operator and the semantics of the language was formally captured. The sequel of this section relies on these references, but an in-depth knowledge is not required.

We first note that the vertices of the knowledge graph of Subsection 2.1 should be interpreted as sets: the object type *person* for instance formally corresponds to the set of all persons under consideration, and the type *number* reflects the set of all numbers. Further, edges should be interpreted as functions: the object type relation *lives on* for instance should be interpreted as the function that assigns to every person the address he or she lives on. Functions therefore have a domain and a codomain: these are the sets associated with the head and the tail vertex of an edge, respectively.

Functions can be combined in several ways to form other useful functions. In Figure 5 we list these combinations; these form the constructs of our language. It should be noted that they form a typed language. By the type of a function we mean the combination of its domain and codomain. Thus, if we write $v : p \rightarrow x$ then p is the domain of v, x is its codomain and $p \rightarrow x$ is its type. Thus, in its basic form, the type of a function corresponds with an edge in the knowledge graph. By a typed language we mean that the application of an operator to its arguments is restricted by conditions that are put on these arguments, or on the types of the arguments. This explains the notation of Figure 5: above the horizontal lines these conditions are written; below the horizontal lines the application of the operator to its arguments is written, as well as the resulting type. Thus, for example, composition is defined solely for two functions, of which the domain of the first equals the codomain of the second, and product is defined for at least two functions that have a common domain.

The semantics of the operators in Figure 5 is as follows, but see Gelsema (2012, 2019) for a thorough exposition. Note that we do not consider coproducts Gelsema (2012) here. Composition is functional composition defined by $(v \circ w)(e) = v(w(e))$ for $e \in q$. The product function $\langle v_1, . . ., v_k\rangle$ maps a $d \in p$ to the tuple $\langle v_1(d), . . ., v_k(d)\rangle$. The projection function $\pi_j(x_1, . . ., x_k)$ maps a tuple $\langle a_1, . . ., a_k\rangle$ with $a_i \in x_i$ to a_j. Suppose that $+$ is defined on x. Then $\alpha(v, w)$ maps an $e \in q$ to the sum $\sum_{d \in w^{-1}(e)} v(d)$, wher it is understood that this sum equals 0 if w^{-1}(e) is the empty set. Let $\sigma(v_1 \sim w_1, . . ., v_k \sim w_k)$ be the set $\{d \in p | v_1(d) = w_1(d), . . ., v_k(d) = w_k(d)\}$. Then $\iota(v_1 \sim w_1, . . ., v_k \sim w_k)$ is the injection of this set into p. Let $\rho(v)$ be the set $\{e \in q | v(d) = e$ for some $d \in p$. Then $\kappa(v)$ is the injection of this set into q.

The intuitive meaning of the constructs of Figure 5 is as follows, where we try to translate them into database terms as much as possible. Roughly, composition corresponds to a `JOIN` clause of a SQL statement. Product is the construct that allows the formation of a database table from its columns. Projection corresponds to a `SELECT` clause. Aggregation can be seen as the combination of a `SUM` with a `GROUP BY` clause in SQL.

composition $$\frac{v : p \rightarrow x,\ w : q \rightarrow p}{v \circ w : q \rightarrow x}$$

product $$\frac{v_i : p \rightarrow x_i,\ k > 1,\ 1 \leq i \leq k}{\langle v_1, \ldots, v_k \rangle : p \rightarrow x_1 \times \cdots \times x_k}$$

projection $$\frac{k > 1,\ 1 \leq i \leq k}{\pi_i(x_1, \ldots, x_k) : x_1 \times \cdots \times x_k \rightarrow x_i}$$

aggregation $$\frac{v : p \rightarrow x,\ w : p \rightarrow q}{\alpha(v, w) : q \rightarrow x}$$

selection $$\frac{v_i, w_i : p \rightarrow x,\ k > 0,\ 1 \leq i \leq k}{\iota(v_1 \sim w_1, \ldots, v_k \sim w_k) : \sigma(v_1 \sim w_1, \ldots, v_k \sim w_k) \rightarrow p}$$

range $$\frac{v : p \rightarrow q}{\kappa(v) : \rho(v) \rightarrow q}$$

Fig. 5. A typed formal language for interpreting requests for statistics.

The selection operator mimics a `WHERE` clause. Finally, the range operator more or less acts as a `DISTINCT` clause.

Now we stress that the operators listed in Figure 7 are an integral part of the knowledge graph, in the sense that they form a derivation mechanism for completing the graph with, for example, arrows that reflect aggregated data. It is in this sense that our approach differs from RDF-like approaches. For instance, it may be obvious to describe a variable like *income* as a triple with subject *person* (or *household*) and with object *number*, as one would perhaps do when constructing an ontology for a statistical domain of interest. It is less obvious however to do the same for the variable *average income*: what is the subject here? In Gelsema (2012) we chose the subject *class of persons* for the latter, which we gave the semantics of a powerset. Also in Gelsema (2012), we proved that this is the right interpretation, in the sense that a homomorphism exists between the knowledge graph and the (aggregate) data it describes. It is exactly this kind of formality we feel RDF-like approaches lack, but it is invaluable when the goal is to use the graph to calculate statistics, like we do in our approach.

All constructs of Figure 5 are described in Gelsema (2019), except for the range operator. During the development of Farseer we discovered its usefulness and its independence from the other operators. We believe that the types and elements approach of Gelsema (2019) can be used to capture the semantics of the range operator as well.

We present some examples of the application of the operators of Figure 5 to the knowledge graph of Subsection 2.1. First, the reader can verify that the set of people that live in The Hague is formalized by the expression

$$\sigma(\mathit{munic.\ name} \circ \mathit{in} \circ \mathit{liveson} \sim \mathit{'s\text{-}Gravenhage} \circ \mathit{all}(\mathit{person})),$$

where *all*(*person*) is the unique function (called *all* in the knowledge graph of Figure 3) with domain *person* and with codomain 1, and where we use that $\circ$ is associative. Note that the composition of *all*(*person*) with the constant *'s-Gravenhage* ‘factors through’ the type 1.

In much the same way, jobs with job title *oogarts* are selected from the full set of jobs by the selection

$$\iota(\mathit{job\ title} \sim \mathit{oogarts} \circ \mathit{all}(\mathit{job})).$$

Note that the types of *job title* and *oogarts* $\circ$ *all*(*job*) coincide, as required by selection. Also note that the type of the inclusion function is

$$\sigma(\mathit{job\ title} \sim \mathit{oogarts} \circ \mathit{all}(\mathit{job})) \rightarrow \mathit{job}.$$

Now suppose that we are interested in persons that have at least one job as an ophthalmologist. This is taken care of by the range operator:

$$\kappa(\mathit{employee} \circ \iota(\mathit{job\ title} \sim \mathit{oogarts} \circ \mathit{all}(\mathit{job}))),$$

which yields a function that maps these ophthalmologists to the full set of persons.

Next, our example query ‘Hoeveel mensen wonen er in Den Haag’ is expressed by the formula

$$\begin{gathered}\alpha(\mathit{one}(\mathit{person}) \circ \iota(\mathit{munic.\ name} \circ \mathit{in} \circ \mathit{lives\ on} \sim \mathit{'s\text{-}Gravenhage} \circ \mathit{all}(\mathit{person})),\\ \mathit{all}(\mathit{person}) \circ \iota(\mathit{munic.\ name} \circ \mathit{in} \circ \mathit{lives\ on} \sim \mathit{'s\text{-}Gravenhage} \circ \mathit{all}(\mathit{person}))).\end{gathered}$$

It can be seen that this expression yields a function of type 1 → *number* and thus denotes a numerical constant, as required.

Dimensional figures come into play when, for example, we require 'Hoeveel mensen wonen in Den Haag naar geslacht?' ('How many people live in The Hague, by sex?'). This is a request for two numbers: the number of males in The Hague and the number of females. It is expressed formally by

$$\begin{gathered}\alpha(\mathit{one}(\mathit{person}) \circ \iota(\mathit{munic.\ name} \circ \mathit{in} \circ \mathit{lives\ on} \sim \mathit{'s\text{-}Gravenhage} \circ \mathit{all}(\mathit{person})),\\ \mathit{sex} \circ \iota(\mathit{munic.\ name} \circ \mathit{in} \circ \mathit{lives\ on} \sim \mathit{'s\text{-}Gravenhage} \circ \mathit{all}(\mathit{person}))),\end{gathered}$$

and yields a function of type *gender* → *number*.

Finally, aggregating over a numerical variable is easy; the request 'Wat is het totale inkomen van Den Haag?' ('What is the total income of The Hague') is expressed by

$$\begin{gathered}\alpha(\mathit{income} \circ \iota(\mathit{munic.\ name} \circ \mathit{in} \circ \mathit{lives\ on} \sim \mathit{'s\text{-}Gravenhage} \circ \mathit{all}(\mathit{person})),\\ \mathit{all}(\mathit{person}) \circ \iota(\mathit{munic.\ name} \circ \mathit{in} \circ \mathit{lives\ on} \sim \mathit{'s\text{-}Gravenhage} \circ \mathit{all}(\mathit{person}))).\end{gathered}$$

and this expression can easily be rewritten in some dimensional form.

Having read the example expressions above, the reader might get a feeling for the way formulas are automatically generated given a request, the lists of tokens, objects and keywords associated with the request, and the class that is estimated for the request. We give an overview of the algorithm that computes such formulas without giving all the details.

During semantic analysis, the first object that is derived from the three lists is the so-called pivot: it is the object type in the knowledge graph from which all other objects in the object list can be reached through paths originating from it. Thus, a request such as 'Hoeveel mensen zijn er?' ('How many people are there?') has *person* as a pivot, but a request such as 'Hoeveel oogartsen zijn er?' ('How many ophthalmologists are there?') has *job* as a pivot.

Note therefore that the pivot might not be the object type that is central in the aggregation process; by this we mean the object type of the variable that is the subject of counting (either *one* or some other numerical variable). The object type that is central in aggregation we call the target. To derive the target from the lists of tokens, objects and keywords, we use a ML algorithm similar to the one explained in Subsection 2.3, trained on the same set of handwritten requests translated to lists of keywords. Again, a Conv1D layer is used as main layer; the output of the model however is now an index into the lists of tokens, objects and keywords that points to the estimated target. We omit further details. Thus, both requests 'Hoeveel mensen zijn er?' and 'Hoeveel oogartsen zijn er?' have *person* as a target.

If pivot and target do not coincide, we look for a path in the knowledge graph originating from the pivot that connects them, and we prefer the shortest if there are more than one. In the case of 'Hoeveel oogartsen zijn er?', this path is simply the object type relation *employee*. This gives an indication for the application of the range operator since, for example, the (inclusion associated with the) set of ophthalmologists is given by

$$\kappa(\mathit{employee} \circ \iota(\mathit{job\ title} \sim \mathit{oogarts} \circ \mathit{all}(\mathit{job})))$$

and thus the total number of ophthalmologists is given by the expression

$$\alpha(one(person) \circ \kappa(employee \circ \iota(job\ title \sim oogarts \circ all(job))),$$
$$all(person) \circ \kappa(employee \circ \iota(job\ title \sim oogarts \circ all(job)))).$$

Next, we note that a selection condition, that is, a pair of arguments to the selection operator, is given by a combination of a variable (*job title* in the expression above) and a constant (*oogarts*) having the same codomain. This means that each constant in the list of objects is an indication of a selection condition; identifying in advance that, for example, *job title* is the preferred variable for the constant *oogarts*, variables are easily found matching constants. Using the knowledge graph, a path from the pivot to the domain of the variable is then sought. If multiple paths exist, then objects in the object list are used as a heuristic, much like pebbles, to find the best matching path. A path is then converted into a sequence of compositions and, together with the corresponding constant variable pair, is given to the selection operator as an argument.

Potential dimensions are extracted from the list of keywords, using certain combinations of keywords as a match, such as $<per>$, $<prep><all>$ and $<prep><ot>$ (see Table 2). We skip the details, but the result of this procedure is a list of indices pointing to objects in the object list that are treated as a dimension. For instance, the result of applying this procedure to the request ‘Aantal inwoners per geslacht en gemeente’ (‘Number of inhabitants per sex and municipality’) is a list of indices pointing to the objects ‘geslacht’ (*sex*) and ‘gemeente’ (*municipality*). Then, in the usual case, appropriate paths are sought from the target to these objects, which are then turned into sequences of compositions, in much the same way as described in the previous paragraph. If there are multiple dimensions, then these compositions are put together inside a product construction, as in

$$\alpha(one(person), \langle munic.\ name \circ in \circ lives\ on,\ sex\rangle),$$

which is the formal representation of ‘Aantal inwoners per geslacht en gemeente’. It yields a function of type $munic.\ names \times gender \rightarrow number$, as required. There are exceptions, but we will leave them out here.

Finally, the class associated with a request, as outlined in Subsection 2.3, gives an indication for the outer operators in its formal representation. For example, a request of class 3 yields a formula that is of the simple form

$$\alpha(v, w),$$

where v and w have a common domain, as required. A request of class 4 has as a general formula

$$(/) \circ \langle \alpha(v, w), \alpha(one(p), w)\rangle,$$

where we assume a division function (/) of type $number \times number \rightarrow number$, and where the domains of v and w equal p. Finally, requests for class 5 are treated with the general formula

$$(/) \circ \langle \alpha(one(p), w), \alpha(one(q), z)\rangle,$$

where w and z are required to have a common codomain. We skip the details for the rest of the classes, but mention only that our formal language has no constructs for ordering. This means that formulas for class 6,. . ., 11 requests must be accompanied with a proper ordering condition.

On the technical side, the semantic analysis part of Farseer was developed with the aid of two Python modules. The first is used for defining object types, variables, constants etcetera, in short: all the colored vertices and edges a knowledge graph is constructed of. Per January 2022, is is now possible to store such a knowledge graph in a graph database, to take advantage of advanced graph searching algorithms. The second Python module is used to store expressions of the kind encountered in this section. An expression is thus a Python object, that can be inspected to retrieve its outer operator, see Figure 5 and its subexpressions. At the lowest level, subexpressions are objects from the first Python module, that is, the primitives of an expression are the vertices and edges of a knowledge graph. The combination of these modules can be seen as a way to infer statistical knowledge from a graph, where the second Python module provides the inference mechanism.

2.5. *Query Generation*

Essentially, the task at hand in the query generation stage is what is usually done by a compiler Aho et al. (1986): the translation of a program written in a source language to a program written in a target language. Here, the source language is the formal language explained in Subsection 2.4 and the target language is one of two common variants of the Structured Query Language (SQL), that is, MySQL Axmark and Widenius (2021) and Transact-SQL Ben-Gan and Moreau (2000). Hence, we assume that the target system on which results for a given request are to be computed is a database management system (DBMS) and it is therefore also assumed that all relevant microdata are stored there. Since our inputs to the compilation procedure consist of formulas that are derived using the type system of Figure 5 and therefore assumed to be syntactically correct, there is no need for lexical analysis or parsing, which are some stages commonly found in compilation routines.

The inductive nature of the formal language depicted in Figure 5 calls for an obvious strategy for compiling a formula of the language, that is, by recursion. Note that each of the constructs of Figure 5 consists of an operator that is applied to a subformula – or a series of subformulas – already assumed part of the language. This means that they can be viewed, roughly but sufficiently, as an algebraic structure in the sense of universal algebra Meinke and Tucker (1992), that is, as a set of operators that can be applied to a number of arguments of some sort.

By inspecting the outer operator of a given formula, its compilation is diverted to a procedure that handles the specifics of the operator. Thus, for each operator a separate procedure is considered and the first step of each is to recursively compile the arguments to the operator (which is sometimes called head recursion).

Key to the functioning of the compilation procedures are the data structures upon which they operate. More specifically but roughly still, the goal of each procedure applied to a given formula is to compute a tuple $t = (s_d, s_c, f, j, w, g, d)$, where

- s_d and s_c are lists of database table column names corresponding to the domain and codomain of the formula, respectively. These column names are meant to fill in the '`SELECT. . .`' part of the resulting SQL query;

- *f* is a database table name meant to supplement the ‘FROM. . .’ part of the resulting query;
- *j* is a list of pairs of database table names and join conditions meant to appear in the ‘JOIN. . . ON. . .’ parts of the resulting query;
- *w* is a list of clauses of the form ‘lhs = rhs’ meant to fill in the ‘WHERE. . .’ part of the resulting query (and each clause is intended to be separated by the ‘AND’ SQL keyword);
- *g* is a list of database table column names meant to fill in the ‘GROUP BY. . .’ part of the resulting query, and
- *d* is a Boolean value indicating whether or not the ‘SELECT. . .’ part should be accompanied with the ‘DISTINCT’ keyword.

Thus, for example, to compile a formula whose outer operator is the composition operator, the objective of the corresponding procedure is to construct a tuple t of the form above, given two other tuples t_l and t_r obtained by recursion, that is, one for the left hand side argument of the composition and one for its right hand side argument. We leave out the specifics of this procedure, as well as the rest of the compilation procedures. We stress that tuples of the given form constitute a wide range of queries, but a subset still of all queries that can be formulated using MySQL or Transact-SQL. They are however all we need for the moment.

Using the approach outlined above, some general technical details need to be considered though. First, table names and table column names should be aliased properly, since there can be several JOIN clauses a table name can appear in (there can be an unbounded number of JOINs in principle). Proper aliases can be derived through concatenation, essentially following consecutive directed edges in corresponding paths in the knowledge graph. Second, we hypothesized in advance that nested queries should be avoided completely, as they may lead to unacceptable execution times. Instead, a potentially nested query (or rather, a proto-query in the form of a tuple t) qualifies to be broken up into several queries to be executed iteratively, and this means that their intermediate results need to be stored in temporary tables. Various approaches can be chosen to decide when to introduce an intermediate query (or to ‘freeze’ a query, as it turned out to be called). At the moment, a reasonable and satisfying approach seems to be to ‘freeze’ a query, once d becomes true or once g becomes nonempty.

3. Results

Of the 2,155 sample requests used for training the classification procedure, only six were misclassified (0.28%) leading to an incorrect semantics. The target index was estimated incorrectly for eleven requests (0.51%) which in all but three cases led to an incorrect formula, or no computed formula at all. In a total of 42 cases (1.95%), the output of the semantic analysis procedure resulted in no formula being computed. In 18 of these cases, this was actually the correct response because answering those requests proved to be beyond the capabilities of the information model. Whenever a formula was produced, the SQL query corresponding to the formula executed with no errors and had the correct semantics relative to the formula. A total of 62 requests (2.88%) produced a formula different from other requests in the same chapter and therefore received ambiguous or incorrect intended semantics. All SQL queries were executed on a small database with fictitious data and those corresponding to the intended semantics of a request produced the correct answer.

One of the main virtues of the Farseer framework, we claim, is its ability to operate across various statistical domains. One constructs a knowledge graph of a particular domain, updating the lookup table from Figure 4 as one goes along, and the framework will adapt to the concepts and vocabulary of that domain, without the need for retraining the framework for request classification or rewriting the routines for interpreting a request. Knowledge graphs and lookup tables need to be built from scratch though, and it takes some experimenting before the lookup table is aligned with the preferred vocabulary of a user. One other desired virtue is its performance: the quick response to a request irrespective of the size of the data on which it operates.

In order to test these hypotheses (i.e., Farseer is domain independent and fast), a data set containing microdata of police-registered crime was selected for experimentation. This dataset contained information of over 900,000 criminal offenses committed in 2016 in the Netherlands, with details about the type of the offense, its location (up to postal code level of detail), the date committed (both start and end dates), the number of suspects involved, as well as other details of the crime, such as the nature of its location (e.g., business premises, terrain, park). From these, the following variables were selected: municipality of the location of the crime, crime type according to criminal law, start date, and the number of suspects. This information was subsequently captured in the knowledge graph of Figure 6 below.

As with the knowledge graph of Figure 3 the knowledge graph of Figure 6 is an English translation of the actual graph used, except for the brown vertices representing constants: of these, *vermogensmisdrijf* translates to *property crime* and *diefstal van fiets* to *bicycle theft*. There are actually four variables associated with crime type, from *crime type level 0* to *crime type level 3*. This is due to the fact that crime type is measured as a hierarchy consisting of four levels, and we have, for example, that *bicycle theft* is a particular form of

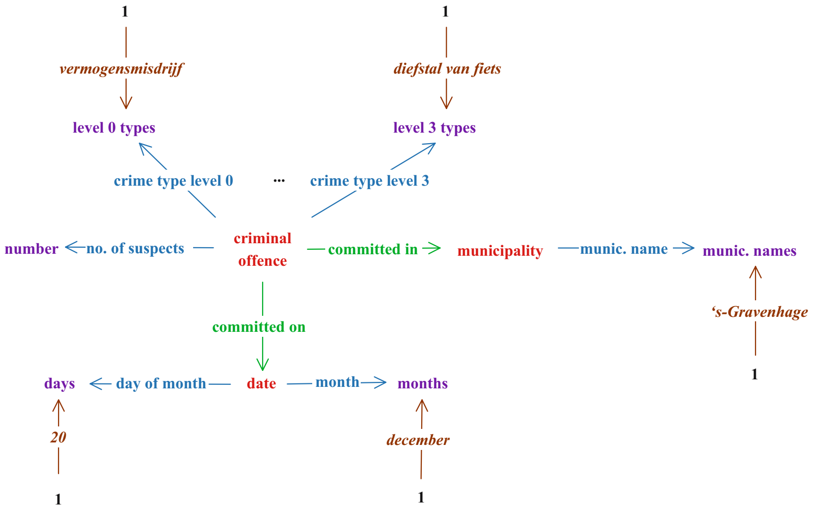


Fig. 6. Knowledge graph of criminal offenses.

property crime. Note that many of the details of Figure 6 are left out: in reality there are almost 400 municipalities (in 2016 in The Netherlands) and almost 100 types of crime at the finest level of detail.

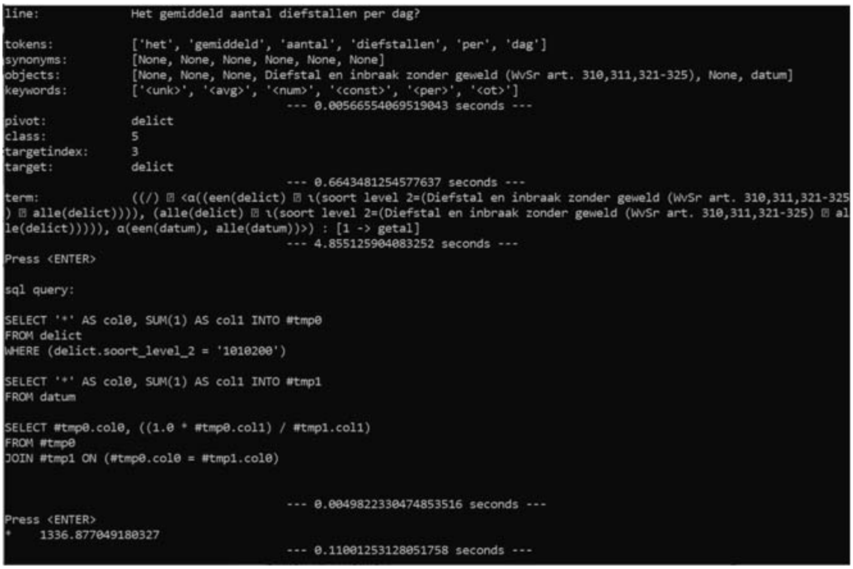
For types of crime in particular, the lookup table explained in Subsection 2.2 was updated. Special attention was given to objects that are characterized by two or more tokens, such as *diefstal van fiets*: in this case, if the combination of the terms *diefstal* (*theft*) and *fiets* (*bicycle*) – or synonyms of them – occurs in a request, then the token that corresponds to the first occurrence of the two is mapped to the constant *diefstal van fiets*. The word embeddings mechanism to detect synonyms was left untouched.

In advance, a collection of 44 requests (see Appendix Section 6) pertaining to criminal offenses and covering all classes but class 6 was assembled. We note that, in particular, the ML method for classifying requests explained in Subsection 2.3 was not updated, so it was hypothesized that the classification mechanism that was learned during development would be sufficient for classifying requests from a new domain of interest. Class 6 requests do not apply here, since there is no obvious concept of 'greatest' criminal offense.

Finally, the microdata containing the criminal offenses was stored in a SQL database, covering three tables: one for offenses, one for days of the year and one for municipalities. The database was not fully normalized. For instance, a table containing the months of the year was lacking; a choice that had unfortunate consequences, since requests that rely on counting months could not be answered. In other words: normalization is preferred, when the counting of some concept is required.

All 44 requests were run on the framework and all were interpreted correctly. Also, the execution of the query that was produced gave the correct result in all cases; this was checked by a visual inspection of the query.

In Figure 7 the outputs of several stages of the Farseer framework are shown, together with their processing times. For example, it is shown that the request is a class 5 request,



```
line:           Het gemiddeld aantal diefstallen per dag?

tokens:         ['het', 'gemiddeld', 'aantal', 'diefstallen', 'per', 'dag']
synonyms:       [None, None, None, None, None, None]
objects:        [None, None, None, Diefstal en inbraak zonder geweld (WvSr art. 310,311,321-325), None, datum]
keywords:       ['<unk>', '<avg>', '<num>', '<const>', '<per>', '<ot>']
                                            --- 0.00566554069519043 seconds ---
pivot:          delict
class:          5
targetindex:    3
target:         delict
                                            --- 0.6643481254577637 seconds ---
term:           ((/) ▨ <α((een(delict) ▨ ι(soort level 2=(Diefstal en inbraak zonder geweld (WvSr art. 310,311,321-325
) ▨ alle(delict)))), (alle(delict) ▨ ι(soort level 2=(Diefstal en inbraak zonder geweld (WvSr art. 310,311,321-325) ▨ al
le(delict))))), α(een(datum), alle(datum))>) : [1 -> getal]
                                            --- 4.855125904083252 seconds ---
Press <ENTER>

sql query:

SELECT '*' AS col0, SUM(1) AS col1 INTO #tmp0
FROM delict
WHERE (delict.soort_level_2 = '1010200')

SELECT '*' AS col0, SUM(1) AS col1 INTO #tmp1
FROM datum

SELECT #tmp0.col0, ((1.0 * #tmp0.col1) / #tmp1.col1)
FROM #tmp0
JOIN #tmp1 ON (#tmp0.col0 = #tmp1.col0)


                                            --- 0.0049822330474853516 seconds ---
Press <ENTER>
*    1336.877049180327
                                            --- 0.11001253128051758 seconds ---
```

Fig. 7. Example of the console output of the Farseer framework.

Table 3. Average and maximum processing times for various stages.

Stage	Average	Maximum
Tokenization and NER	0.014	0.055
Request classification	1.896	10.780
Interpretation	4.337	13.515
Query generation	0.006	0.110
Query execution	0.130	0.538

that it took nearly five seconds to generate a formula for it, that the time to translate the formula into a query took less than 0.005 seconds, and that the execution time of the query was about 0.1 seconds. The top two time indications reflect the time it took to perform tokenization and NER, and to perform request classification, respectively. The final output of the execution of the generated query is given at the bottom: apparently there were about 1,336 cases of theft in 2016 in The Netherlands on an average day. The request translates to *the average number of cases of theft per day*.

In Table 3, the average processing times in seconds as well as the maximum processing times for each of the stages is shown: these were collected from the runs of the 44 requests.

As can be seen, request classification and the generation of a formula take average execution times of nearly two seconds and over four seconds, respectively, and these stages take up most of the overall processing time by far.

4. Discussion

The authors are well aware that the way the experiment of Section 3 was designed and conducted can be criticized for many reasons. Ideally, the 44 requests (or preferably more) should have been written by a body with no prior knowledge of the capabilities of the framework and with no expert knowledge of the domain of crime. In this way the casual language and vocabulary of a typical end-user can be simulated without biases towards the use of expert language, or towards the formulation of requests known in advance to be interpreted correctly with a high probability. Then, an independent second body (a crime domain expert preferably) should have been given the task of manually interpreting these requests prior to the experiment and calculating their results wherever possible, independent of the framework. Thus, a list of interpretations and results that serve the validation of the framework should have been available prior to conducting the experiment. Only then the requests should have been run by the framework, and the results compared with this independent list.

As a measure of performance of Farseer in such an experiment in terms of information quality, we consider the use of the InfoQ Kennet and Shmueli (2016) frameworks' dimensions, of operationalization and communication in particular.

The 100 percent score for the correct interpretation of the 44 requests indicates that there is such a bias. This means that at this stage we do not know how well the framework performs in the translation of casual language to expert language. Indeed, the framework breaks down when confronted with more exotically formulated requests such as *Hoe vaak is er iets in de fik gestoken?* (which translates roughly to: *How many times was something torched?*). In this example, the framework is not able to make a connection between the word *fik* (*torched*) and the concept of *brandstichting* (*arson*), although *fik* is included in the vocabulary of 150,000 words taken from Wikipedia. It seems that this is due to the somewhat limited capabilities of the word embeddings technique to discover synonyms: as mentioned, word embeddings record similarities of words (in some context) and they include, for example, antonyms as well. As a possible improvement of the framework, we consider the use of thesauri. Also, experiments would benefit from the use of several language models, including casual and expert. We note that the word *fik* has also another meaning in Dutch, that is, finger.

From Table 3 it can be concluded that query generation appears to be fast and that query execution has a reasonable processing time on average. Also, tokenization and NER do not take much time to complete, as can be expected, although tokenization still is implemented in an ad hoc fashion. However, the time (both average and maximum) it takes to perform interpretation we think is inadmissible for a framework that advocates an on-the-fly response. One probable reason for this poor performance is that interpretation depends heavily on finding (optimal) paths between two given vertices in the knowledge graph. As of this moment however, the knowledge graph is implemented as a set of Python objects stored in memory and, more importantly, the searching algorithms that operate on it are implemented naively. We expect a much better performance when the knowledge graph is stored using a graph database of some sort, taking advantage of optimized graph searching engines. As of January 2022, a graph database was implemented and achieved a tenfold decrease in the time needed to do semantic analysis.

We also believe that the time it takes to classify requests is still too high. Possibly in this stage also, using a graph database will reduce computation time. Moreover, by the Keras, Chollet (2018) (and TensorFlow, Géron 2017) deep learning package that we employ, it is recommended using GPUs instead of CPUs to speed up computation, both for training and inference; we have been employing CPUs exclusively so far. Therefore we believe that also the performance of the classification stage can be improved considerably.

The reader may have noticed from the formulation of the requests in Appendix that the word ‘total’ is used more often than seemingly necessary. For instance, one could argue that the use of ‘total’ is superfluous in the correct interpretation of requests nos. 15 up to and including 18. However, at the moment the framework will, rather autistically, treat those as requests for microdata if ‘total’ is left out. For instance, in the case of number 15, the framework will return all cases of arson (i.e., identifiers for these cases) together with the number of suspects for each individual case. An obvious remedy for the overuse of the word ‘total’ is to exclude requests for microdata altogether, that is, divert requests for the classes handling microdata (classes 1 and 6 in particular) to their aggregate counterparts, wherever possible. Of course, in order for the framework to be employed in handling requests from outside the office, such should be the case to begin with.

The reader may also wonder if the use of an intermediate formal language is strictly necessary: why not immediately output a SQL query as the result of the interpretation stage? While we believe that this is possible in principle, allowing an intermediate formal language is preferable for several reasons. First, the formal language explained in Subsection 2.4 is a language that has the vertices and edges of a knowledge graph as primitives: it is therefore a language built “on top of” the graph constructs. In this way the language stays close to the concepts under consideration (object types, variables, etc.)

which is preferable for the task of semantic analysis. Second, it is a language designed to express meaningful requests (for aggregates or for microdata) only, excluding requests that make no sense: this is the sole purpose of the type system of the language. In contrast, it is not so hard to come up with SQL queries that are nonsensical (given the information captured in the knowledge graph); see, for example, the Discussion in Gelsema (2012). For these reasons, we believe that the intermediate formal language is better suited for interpretation than a SQL-like language. Finally, considering two steps (interpretation and query generation) instead of just one allows us to separate the task of pure interpretation from the task of constructing an efficient query. We believe that the resulting code is therefore easier to understand, maintain and expand.

4.1. Notes on Other Methods

Within the scientific domain of text-to-SQL systems (see Katsogiannis-Meimarakis and Koutrika (2021) for an overview), Farseer adopts mostly traditional rule-based techniques to translate a request to SQL. The combined use of an intermediate language, a graph for describing the domain, and some sort of lookup table to map tokens to elements of the graph is also seen in ATHENA, Saha et al. (2016). In both approaches, once a formula or expression in the intermediate language is produced, a syntactically valid SQL query can be derived from it. However, we feel that since our language is more strongly typed, it also aids in the generation of queries that are semantically valid. Notably, ATHENA does not use ML techniques.

Neural network approaches to the text-to-SQL problem, according to Katsogiannis-Meimarakis and Koutrika (2021), fall into three different categories. First sequence-to-sequence approaches produce sequences of SQL tokens together with database schema elements, of which Seq2SQL, Zhong et al. (2017) is an example. Sequence-to-sequence approaches however are prone to syntactic errors, which we avoid by employing our intermediate formal language. Second, grammar-based approaches generate grammar rules, which, when applied, produce a SQL query. This dramatically reduces the amount of ill-formed SQL queries. RAT-SQL, Wang et al. (2020) is a notable grammar-based text-to-SQL system, as is IRNet, Guo et al. (2019)). Interestingly, the latter also uses an intermediate language from which to generate SQL. Third, sketch-based slot-filling approaches aim at predicting certain parts of a SQL statement, turning the SQL generation problem into a classification problem, as we do. An example is SQLNet, Xu et al. (2017). As a downside, this type of text-to-SQL systems is unsuited for generating complex queries.

Of course, one of the main issues in ML approaches to text-to-SQL problems is the need for large amounts of training data, which typically consist of requestquery pairs and which must be curated manually. Our approach partly avoids this problem: only manually categorized requests need to be used for training purposes. On the other hand, DBPal, Weir et al. (2020) is capable of automatically creating synthetic training data in order to improve existing text-to-SQL systems based purely on ML.

Finally, we feel that our approach best adopts to some practices commonly found in official statistics, such as the use of classification schemes and code lists that separate codes found in data from their meaning addressed in their associated labels. By incorporating these elements from classification schemes in our knowledge graph, we can exploit them in the translation of requests. Also, we feel that the classes of requests we examined, and many more need to be examined, are typical for the kinds of requests users of official statistics need answered.

5. Conclusion

We have presented Farseer: a framework for calculating on-the-fly statistics in response to a user request. We demonstrated that the framework is domain independent: all domain specific information is captured in a knowledge graph that is treated as an input to the framework. We also showed that the framework allows for reasonable query generation and execution times. The time to formally interpret a request still needs to be improved. However, we are currently experimenting with graph databases to exploit their searching capabilities. The first results indicate that we can speed up the interpretation phase of Farseer by a factor of ten.

Further research is needed to give some reliable data on the performance of the framework with respect to the translation of common language to expert language. We plan to do further (this time carefully designed) experiments on a new statistical domain for this purpose in the short term.

The framework still has some other shortcomings; one is its lack of reporting, providing for human readable explanations on how a request was interpreted. We feel that here also ML might prove useful, as reporting can be seen as an instance of a translation problem. After all, all the information necessary to produce a report is captured by the formula that is the result of the interpretation stage. Thus, generating a report requires translating a formula to a human readable form. How to apply ML in this case, especially given the fact that no data is available for training in a supervised learning mode at the moment, still requires some research though.

During the development of Farseer much attention was given to the semantic analysis and query generation stages, as we feel that these constitute our main contributions to the field of NLP applied to statistics. Consequently (and intentionally) less time was spent on, for example, tokenization and NER, for which we knew in advance that much more sophisticated algorithms exist than the ones we implemented. Currently we are exploring the use of spaCy, Vasiliev (2020): a Python package that implements advanced techniques for many NLP tasks. We believe that this gives ample room for improving the NLP tasks at the beginning of our pipeline.

One major obstacle for implementing the framework for the purpose of responding to end-user requests (i.e., requests from outside the office) is disclosure control, or rather the lack of it for real time, on-the-fly applications such as ours. We hope that this article is an incentive to researchers in the field for developing efficient methods for reliable statistical disclosure control routines that can be put into practice, in order to provide for a fast, reliable and safe response to an arbitrary request for a statistic. We also hope for a discussion about a sensible risk-utility trade-off for applications such as ours.

6. Appendix: Overview of Requests on Criminal Offences

This section lists the 44 requests concerning the domain of criminal offenses considered in Section 3.

Table 4. Requests on the subject of crime.

No.	Request	English translation	Class
1	Totaal aantal verdachten van geweldpleging	Total number of suspects of violence	3
2	Aantal verdachten van geweldpleging	Number of suspects of violence	1
3	Dagen waarop brandstichting is gepleegd	Days in which there were cases of arson	1
4	Maanden waarin brandstichting is gepleegd	Months in which there were cases of arson	1
5	Gemeenten met vernielingen aan auto's	Municipalities with destruction of cars	1
6	Gemeenten met vernielingen aan auto's in december	Municipalities with destruction of cars in December	1
7	Welke dagen zijn er?	What days are there?	1
8	Gevallen van dierenmishandeling met datums	Cases of animal abuse with dates	1
9	Aantal delicten	Number of criminal offenses	2
10	Aantal delicten in Aalsmeer	Number of criminal offenses in Aalsmeer	2
11	Aantal delicten per gemeente	Number of criminal offenses by municipality	2
12	Aantal delicten per dag	Number of criminal offenses by day	2
13	Aantal brandstichtingen per dag en gemeente	Number of cases of arson by day and municipality	2
14	Aantal gevallen van stalking in Amsterdam	Number of cases of stalking in Amsterdam	2
15	Totaal aantal verdachten van brand-stichting	Total number of suspects of arson	3
16	Totaal aantal verdachten van geweld in Leiden	Total number of suspects of violencein Leiden	3
17	Wat is het aantal verdachten in mei in totaal?	What is the number of suspects in May in total?	3
18	Het totaal aantal verdachten in Amsterdam in mei van verkrachting	Total number of suspects in Amsterdam in May of rape	3
19	Het gemiddeld aantal verdachten van geweld	The average number of suspects of violence	4
20	Wat is gemiddeld het aantal verdachten van mishandeling?	What is, on average, the number of suspects of mistreatment?	4
21	Het gemiddelde aantal verdachten van vernieling per gemeente	The average number of suspects of destruction by municipality	4
22	Wat is gemiddeld het aantal verdachten van cybercrime per maand?	What is, on average, the number of suspects of cybercrime by month?*	4

Table 4. Continued

No.	Request	English translation	Class
23	Het gemiddeld aantal verdachten in Delft van inbraak per maand	The average number of suspects in Delft of burglary by month	4
24	Wat is het gemiddeld aantal delicten per dag?	What is the average number of offenses per day?	5
25	Het gemiddeld aantal diefstallen per dag?	The average number of cases of theft per day?	5
26	Wat is het gemiddeld aantal vernieling [sic.] per dag in Delft?	What is the average number of cases of destruction per day in Delft?	5
27	Wat is het gemiddeld aantal brand-stichtingen per gemeente?	What is the average number of cases of arson per municipality?	5
28	Het gemiddeld aantal gevallen van geweld per gemeente in mei	The average number of violence per municipality in May	5
29	Op welke dag zijn er de meeste delicten gepleegd?	On what day were the highest number of offenses committed?	7
30	In welke gemeente zijn er de meeste overvallen gepleegd?	In which municipality were the highest number of robberies?	7
31	Wat is de gemeente met de meeste brandstichtingen?	Which municipality has the most cases of arson?	7
32	Op welke dag zijn er de meeste gevallen van inbraak in Delft gepleegd?	What day has the highest number of cases of burglary in Delft?	7
33	In welke gemeente zijn er de meeste verdachten in totaal?	Which municipality has the highest number of suspects in total?	8
34	Op welke dag waren er het [sic.] meeste verdachten in totaal?	Which day had the highest number ofsuspects in total?	8
35	In welke gemeente zijn er in mei demeeste verdachten in totaal?	Which municipality has in May the highest number of suspects in total?	8
36	Op welke dag in Leiden waren er de meeste verdachten van brandstichting in totaal?	On which day in Leiden were there the highest number of suspects of arson in total?	8
37	In welke gemeente zijn er gemiddeld het meeste aantal verdachten?	Which municipality has, on average, the highest number of suspects?	9
38	Op welke dag in mei zijn er gemiddeld het grootste aantal verdachten?	Which day in May has, on average, the highest number of suspects?	9

Table 4. Continued

No.	Request	English translation	Class
39	De gemeente met het grootste aantal verdachten van brandstichting	The municipality with the highest number of suspects of arson	10
40	De dag met het grootste aantal verdachten van inbraak	The day with the highest number of suspects of burglary	10
41	De dag met het grootste aantal verdachten van inbraak in Leiden	The day with the highest number of suspects of burglary in Leiden	10
42	Welke gemeente heeft gemiddeld het grootst aantal gevallen van inbraak per dag?	Which municipality has, on average, the highest number of cases of burglary per day?	11
43	In welke gemeente wordt gemiddeld het meeste brand gesticht per dag?	Which municipality has, on average, the highest number of cases of arson per day?	11
44	In welke gemeente wordt in totaal het meeste brand gesticht?	Which municipality has, in total, the highest number of cases of arson?	7

*Cybercrime is inaccurately and narrowly interpreted as *computerhuisvredebreuk* which roughly translates as *trespassing by computer*. The data set covering criminal offenses does not record the broader notion of cybercrime as a separate category though.

7. References

Aho, A.V., R. Sethi, and J.D. Ullman. 1986. *Compilers: principles, techniques, and tools*. Boston: Addison-Wesley Longman Publishing Co.

Andrews, P.R. 2002. *An Introduction to Mathematical Logic and Type Theory: To Truth Through Proof*. Dordrecht: Kluwer Academic Publications.

Axmark, D., and D. Widenius. 2021. *MySQL 8.0 Reference Manual*. Redwood Shores: Oracle Corporation. Available at: http://dev.mysql.com/doc/refman/8.0/en/ (accessed January 2023).

Barendregt, H.P. 1984. *The Lambda Calculus; Its Syntax and Semantics*. Amsterdam: Elsevier Science B.V.

Ben-Gan, I., and T. Moreau. 2000. *Advanced Transact-SQL for SQL Server 2000*. New York: Springer-Verlag.

Bojanowski, P., E. Grave, A. Joulin, and T. Mikolov. 2017. "Enriching Word Vectors with Subword Information." *Transactions of the Association for Computational Linguistics* 5: 135–146. DOI: https://doi.org/10.1162/tacl_a_00051.

Brickley, D., and R.V. Guha. 2014. *RDF Schema 1.1*. Massachusetts: W3C. Available at: http://www.w3.org/TR/2014/REC-rdf-schema-20140225/ (accessed January 2023).

Chollet, F. 2018. *Deep Learning with Python*. Shelter Island: Manning Publications Co.

Codd, E.F. 1970. "A Relational Model of Data for Large Shared Data Banks." *Communications of the ACM* 13: 377–387. DOI: https://doi.org/10.1145/362384.362685.

Enderle, T., S. Giessing, and R. Tent. 2006. "Designing Confidentiality on the Fly Methodology – Three Aspects." *Proceedings of PSD* LNCS 11126: 28–42. DOI: https://doi.org/10.1007/978-3-319-99771-1_3.

Fraser B, and J. Wooton. 2006. "A proposed method for confidentialising tabular output to protect against differencing." *Monographs of Official Statistics*. Work session on Statistical Data Confidentiality: 299–302. Luxembourg: Eurostat-Office for Official Publications of the European Communities. Corpus ID: 53573926.

Gelsema, T. 2008. "General requirements for the soundness of metadata models." Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS). Available at: https://www.researchgate.net/publication/334708026_General_requirements_for_the_soundness_of_metadata_models (accessed January 2023).

Gelsema, T. 2012. "The Organization of Information in a Statistical Office." *Journal of Official Statistics* 28(3): 413–440. Available at: https://www.scb.se/contentassets/-ca21efb41fee47d293bbee5bf7be7fb3/the-organization-of-information-in-a-statistical-office.pdf.

Gelsema, T. 2019. "The Logic of Aggregated Data." *Acta Cybernetica* 24(2): 211–248. DOI: https://doi.org/10.14232/actacyb.24.2.2019.4.

Ǵeron, A. 2017. *Hands-On Machine Learning with Scikit-Learn & Tensorflow*. Sebastopol: O'Reilly Media Inc.

Guo, J., Z. Zhan, Y. Xiao, J.G. Lou, T. Liu and D. Zhang. 2019. "Towards Complex Text-to-SQL in Cross-Domain Database with Intermediate Representation." DOI: https://doi.org/10.48550/arXiv.1905.08205.

Hayes, P.J., and P.F. Patel-Schneider. 2014. *RDF 1.1 Semantics.* Massachusetts: W3C. Available at: http://www.w3.org/TR/2014/REC-rdf11-mt-20140225/ (accessed January 2023).

Indurkhya, N., and F.J. Damerau. 2010. *Handbook of Natural Language Processing, Second Edition*. Boca Raton: Chapman & Hall/CRC.

Ji, Z., Z.C. Lipton and C. Elkan. 2014. "Differential privacy and machine learning: a survey and review." DOI: https://doi.org/10.48550/arXiv.1412.7584.

Katsogiannis-Meimarakis, G., and G. Koutrika. 2021. "Deep Learning Approaches for Text-to-SQL Systems." In Proceedings of the 24th International Conference on Extending Database Technology (EDBT), March 23–26: 710– 713. Nicosia. Available at: https://openproceedings.org/2021/conf/edbt/p306.pdf (accessed January 2023).

Kennet, R.S., and G. Shmueli. 2016. "From Quality to Information Quality in Official Statistics." *Journal of Official Statistics* 32(4): 867–885. DOI: https://doi.org/10.1515/-jos-2016-0045.

Marley, J.K., and V.L. Leaver. 2011. "A method for confidentialising user-defined tables: statistical properties and a risk-utility analysis." In Proceedings of 58th World Statistical Congress: International Statistical Institute, Dublin. Available at: https://2011.isiproceedings.org (accessed January 2023).

Meinke, K., and J.V. Tucker. 1992. “Universal Algebra.” In *Handbook of Logic in Computer Science, Vol. I: Background; Mathematical Structures* edited by S. Abramsky, M. Gabbay and T. Maibaum: 189–411. Oxford: Oxford Science Publications.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” DOI: https://doi.org/10.48550/arXiv.1301.3781.

Motik, B., P.F. Patel-Schneider and B. Cuenca Grau. 2012. *OWL 2 Web Ontology Language Direct Semantics (Second Edition)*. Massachusetts: W3C. Available at: http://www.w3.org/TR/2012/REC-owl2-direct-semantics-20121211/ (accessed January 2023).

Saha, D., A. Floratou, K. Sankaranarayanan, U. Farooq Minhas, A.R. Mittal and F. Ozcan. 2016. “ATHENA: An Ontology-Driven System for Natural Language Querying over Relational Data Stores.” In Proceedings of the VLDB Endowment 9(12): 1209–1220. Available at https://vldb.org/pvldb/vol9/p1209-saha.pdf (accessed January 2023).

Thompson, G., S. Broadfoot, and D. Elazar. 2013. “Methodology for the automatic confidentialisation of statistical outputs from remote servers at the Autralian Bureau of Statistics.” Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, October 28–30. Ottawa. Available at: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2013/Topic_1_ABS.pdf (accessed January 2023).

Vasiliev, Y. 2020. *Natural Language Processing with Python and spaCy, A Practical Introduction*. San Francisco: No Starch Press.

Wang, B., R. Shin, X. Liu, O. Polozov, and M. Richardso. 2020. “RAT-SQL: Relation-Aware Schema Encoding, and Linking for Text-to-SQL Parsers.” DOI: https://doi.org/10.48550/arXiv.1911.04942.

Weir, N., P. Utama, A. Galakatos, and A. Crotty. 2020. “DBPal: A Fully Pluggable NL2SQL Training Pipeline.” In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, June 14–19: 2347–2361. New York: Association for Computing Machinery. Available at: https://dl.acm.org/doi/10.1145/3318464.3380589 (accessed January 2023).

Xu, X., C. Liu, and D. Song. 2017. “SQLNet: Generating Structured Queries From Natural Language Without Reinforcement Learning.” DOI: https://doi.org/10.48550/arXiv.1711.04436.

Zhong, V., C. Xiong, and R. Socher. 2017. “Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning.” DOI: https://doi.org/10.48550/arXiv.1709.00103.

Received February 2021
Revised February 2022
Accepted January 2023

Journal of Official Statistics, Vol. 39, No. 3, 2023, pp. 381–398, http://dx.doi.org/10.2478/JOS-2023-0017

Database Reconstruction Is Not So Easy and Is Different from Reidentification

Krishnamurty Muralidhar[1] *and Josep Domingo-Ferrer*[2]

In recent years, it has been claimed that releasing accurate statistical information on a database is likely to allow its complete reconstruction. Differential privacy has been suggested as the appropriate methodology to prevent these attacks. These claims have recently been taken very seriously by the U.S. Census Bureau and led them to adopt differential privacy for releasing U.S. Census data. This in turn has caused consternation among users of the Census data due to the lack of accuracy of the protected outputs. It has also brought legal action against the U.S. Department of Commerce. In this article, we trace the origins of the claim that releasing information on a database automatically makes it vulnerable to being exposed by reconstruction attacks and we show that this claim is, in fact, incorrect. We also show that reconstruction can be averted by properly using traditional statistical disclosure control (SDC) techniques. We further show that the geographic level at which exact counts are released is even more relevant to protection than the actual SDC method employed. Finally, we caution against confusing reconstruction and reidentification: using the quality of reconstruction as a metric of reidentification results in exaggerated reidentification risk figures.

Key words: Database privacy; database reconstruction; statistical disclosure control; differential privacy.

1. Introduction

Database reconstruction seems to be the nemesis of official statistics and statistical data release as they have been known so far. According to the U.S. Census Bureau's Chief Scientist:

> This (Dinur and Nissim's database reconstruction) theorem is the death knell for public-use detailed tabulations and microdata sets as they have been traditionally prepared. (Abowd 2017; Abowd et al. 2019)

Whenever a database contains personal information on a set of respondents, data protection legislation may require the organization in charge of a database, called "controller" in the European legal parlance (GDPR 2016), to take steps to protect respondent privacy. SDC, Statistical disclosure control (Dalenius 1977; Hundepool et al. 2012) is a discipline that provides methods to this end. SDC methods operate by masking, that is, altering, the data to be

[1] University of Oklahoma Price College of Business, Dept. of Marketing and Supply Chain Management, 307 West Brooks, Adams Hall Room 10 Norman, OK 73019, U.S.A. Email: krishm@ou.edu
[2] Universitat Rovira i Virgili, Department of Computer Engineering and Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia Av. Països Catalans 26, 43007 Tarragona, Catalonia. Email: josep.domingo@urv.cat

Acknowledgments: Insightful comments by Constance F. Citro to an earlier version of this article are gratefully acknowledged. Support from the European Commission (projects H2020-871042 "SoBigData++" and H2020-101006879 "MobiDataLab"), the Government of Catalonia (ICREA Academia Prize to J. Domingo-Ferrer and grant no. 2021 SGR 00115), and the Spanish Research Agency (project PID2021-123637NB-I00 "CURLING"). is gratefully acknowledged.

protected; masking can be based on data perturbation, on reduction of detail or even on generating synthetic data that preserve some of the statistics of the original data. Depending on when masking is applied, the SDC literature distinguishes among “local” protection (where data are masked by respondents themselves before being collected), “input” protection (data are masked by the controller after collecting them and all subsequent queries are answered based on the masked data) and “output” protection (queries are computed on the true original respondent data and the query outputs are masked before being released).

The process of forming a database can take place at a certain point in time or be continuous during all the lifetime of the database. In the former case, the database is said to be static, whereas in the latter case it is said to be dynamic. In static databases, the data are first collected and then they are structured to form the database. This yields a “frozen” database which is subsequently used to answer any queries. In contrast, dynamic databases contain data that are periodically changing, with new records or even attributes being added and removed over time. Organizational and corporate databases (e.g., containing data on customers, orders, and so on.) are usually dynamic. Static databases are typical in data collected for research and certain data gathered by government agencies. Obviously, not all government data are static, but in many situations they are.

Output protection is the most convenient option for dynamic databases, as it avoids having to create masked versions of the underlying changing data. Whatever the type of protection, the level of protection achieved depends on the extent to which the data have been modified (Dwork et al. 2006). In general, the greater the modification, the greater the protection but the lesser the accuracy and hence the utility of the protected data. In particular, if the original data contain outliers or unique records, greater modification may be necessary. For further discussions on SDC and the privacy-utility trade-off, (see Traub et al. 1984; Adam and Worthmann 1989; Duncan et al. 2001; Hundepool et al. 2012).

The possibility of mounting reconstruction attacks has been known for decades, and a formal theory of reconstruction attacks was developed by Dinur and Nissim twenty years ago (Dinur and Nissim 2003). U.S. Census Bureau methodologists recently stated that such attacks are no longer just a theoretical possibility, but a practical danger. Hence, they advocate using differential privacy (Dwork et al. 2006; Dwork and Roth 2014) – DP in what follows–to protect the statistical outputs of the U.S. 2020 Decennial Census as a way to thwart reconstruction of the underlying microdata. The decision to use DP motivated a lawsuit from the State of Alabama against the U.S. Department of Commerce, basically arguing utility loss (and a delay in the data release) (Alabama 2021). This lawsuit was backed by 16 other states (Associated Press 2021), but it was recently rejected by the judges (Percival 2021), on the grounds that no damage to Alabama will be provable until the DP Census data are made available.

In April 2021, the U.S. Census Bureau published a version of the 2010 Decennial Census using their new DP-based methodology, called DAS. After studying that version, several users have expressed their concerns about the utility loss caused by DAS (Kenny et al. 2021; Ruggles and Van Riper 2022; Hotz et al. 2022; Dove 2021; Schneider 2022).

Since using a DP-based methodology to prevent reconstruction is controversial in terms of utility, it remains extremely relevant to examine the real danger of reconstruction attacks and the extent to which such attacks can be warded off by DP or other methods at a reasonable utility cost.

1.2. Contribution and Plan of this Article

In this article, we reassess the risk of the original data being reconstructed by an attacker based on the protected query outputs. We first give background on reconstruction attacks (Section 2). We then examine the protection that DP can offer against reconstruction (Section 3). After that (Section 4), we discuss the critical relevance for reconstruction of the geographic level at which exact counts are reported. In Section 5 we highlight the differences between reconstruction and reidentification: using the quality of reconstruction as a measure of reidentification risk results in exaggerated reidentification risk. Conclusions and future research lines are summarized in Section 6.

2. Reconstruction Attacks

Reconstruction attacks have been known for a long time in the literature. In their pioneering work, Denning and Schlorer (1980) showed that a poorly designed database query answering system based on *output perturbation* can easily lead to disclosure of some or even all of the database records. The tool they used was the tracker attack, a carefully crafted sequence of queries aimed at isolating and disclosing specific records.

A more formal analysis of the ability to reconstruct the contents of a database using only the outputs of queries was formulated by Dinur and Nissim (2003), hereafter DN. The main step forward is their discovery that the attacker does not even need to be careful when constructing her sequence of queries. The authors assume the database is an n-bit string, that is, it contains records each of which takes value 0 or 1. They further assume all queries to be of the form “How many records in this subset are 0’s?” or “How many records in this subset are 1’s?”. In their setting, the response to every query is computed as the true query answer plus an error E bounded in an interval ($-B, B$) for some $B > 0$. Thus, it is clear that DN assume that protection of query outputs is performed via *output perturbation* and that the error is strictly bounded.

A database reconstruction, according to DN, is a record-by-record reconstruction of the original values such that the *distance between the reconstructed values and the original values is within specific accuracy bounds*. Thus, DN’s attacks are aimed at inferring the value of *each record in the original database* with a high level of accuracy. They consider two different attackers depending on their computational power:

1. **Exponential attacker.** This attacker is able to issue all possible queries. In practice, such an adversary is only realistic for *small databases*, because, say for $n \geq 100$, it would take years or decades to issue all possible queries, even with the fastest computers available. To protect against such an attacker, the output of any query is modified by adding random noise in $[-B, B]$. If the differences between the query responses obtained on the target original database and the corresponding query responses obtained on a specific candidate database are within B, then the candidate database represents a reconstruction of the original database. DN show that, in this case, the candidate database is within distance $4B$ of the target original database, where both databases are taken as binary n-vectors. Thus, unless the value B is relatively large, the candidate database is a *good* reconstruction of the original database. DN proved that, in order to prevent such a good reconstruction, B must be non-negligible compared to n, that is, B must be $O(n)$.

2. **Polynomial attacker.** This attacker issues a number of queries that is polynomial in n, which is feasible for large databases. The database protects the query outputs by adding random noise in $[-B, B]$. Using these protected query outputs, the attacker solves a linear programming problem to reconstruct the database. DN show that, with high probability, the reconstructed database is close to the original database as long as B is within $o(\sqrt{n})$ where the "little o" notation means much smaller than $\sqrt{n}$ as n grows. Hence, to achieve protection, B must be $O(\sqrt{n})$ (with "big O") for a non-negligible proportion of queries.

Thus, DN conclude that, unless the noise added to query outputs is commensurate to the size of the database ($O(n)$ for an exponential adversary and $O(\sqrt{n})$ for a polynomial adversary), the attacker is able to recreate the database. Is a noise level at least $O(\sqrt{n})$ realistic? Consider a database of size $n = 1{,}000{,}000$. What is being required is that the answers to a *non-negligible proportion* of queries differ from the corresponding true answers by about 1,000. Note that this noise does not have to be applied to all queries. Furthermore, a perturbation of about 1,000 is relatively small compared to the size of the database and to queries that may involve several hundred thousand records. Thus, the noise level required to protect against a polynomial adversary seems affordable in many situations.

Without question, DN give very relevant insights into database reconstruction using only responses to queries. The authors give a theoretical framework that explains the reconstruction risk as a function of the adversary's computational power and the noise applied to query outputs. Yet, *providing a theoretical framework for database reconstruction does not mean that every database can be reconstructed.*

For one thing, *the results by DN apply only to output perturbation, but not to local or input protection*. This is explicitly acknowledged by DN when they mention the "CD Model":

> **The CD Model.** The database algorithm above essentially creates a "private" version of the database d', and then answers queries using d'. Note that a user may retrieve the entire content of d' by querying $q_i = \{i\}$ for $1 \leq i \leq n$, after which she may answer all her other queries by herself. This result indicates that it is in some cases possible to achieve privacy in a "CD model", where users get a "private" version of the database (written on a CD), which they may manipulate (say, without being restricted to statistical queries).

Specifically, if local or input masking are implemented, the responses to all queries are based on the masked database. Hence, *for local or input perturbation, the DN framework can only reconstruct the masked database.* Now, if the local or input masking are configured to adequately protect the original database (e.g., using RR at the respondent's or microdata SDC methods described in Hundepool et al. 2012), reconstructing the masked database should not entail disclosure of sensitive information.

3. The Performance of Differential Privacy Against Reconstruction

In Dwork (2011) and Garfinkel et al. (2019), the purported solution to the reconstruction vulnerability of output-protected data is differential privacy (DP). DP was introduced by Dwork et al. (2006) as a framework for quantifying the disclosure risk associated with answering queries based on a confidential database. Assume an adversary submits a query to the database and obtains a query response R. ϵ-DP requires that, given two databases D and D' that differ in one record, and for all subsets S of the space of query responses

$$\Pr(R \in S|D \text{ is used}) \leq e^{\epsilon} \times \Pr(R \in S|D' \text{ is used}). \tag{1}$$

Essentially, DP requires that, by observing R, it must be indistinguishable within a factor e^{ϵ} whether the database D or the database D' are being used. When $\epsilon = 0$, this requirement implies that the database in use must be completely indistinguishable when observing R. In this case, the value of the record differing between D and D' stays completely confidential in spite of R being returned to the adversary. The value ϵ is usually called "privacy budget" and it should be small for the privacy condition of Expression (1) to be meaningful: Dwork (2011) recommended ϵ to be "say, 0.01, 0.1, or in some cases, ln 2 or ln 3."

A well-known property of DP is sequential composition: if k queries are individually answered with privacy levels $\epsilon_1, \epsilon_2, \ldots, \epsilon_k$, respectively, the extant privacy level after answering all k queries is $\epsilon_1 + \epsilon_2 + \ldots + \epsilon_k$.

A relaxation of DP called (ϵ, δ)-DP has also been proposed and is defined as

$$\Pr(R \in S|D \text{ is used}) \leq e^{\epsilon} \times \Pr(R \in S|D' \text{ is used}) + \delta, \tag{2}$$

where δ is the relaxation parameter. The value of δ is often interpreted to imply that ϵ-DP is satisfied with probability $1-\delta$. But a closer comparison between Expressions (1) and (2) suffices to realize that (ϵ, δ)-DP can hold without ϵ-DP being satisfied for *any* query.

The usual procedure to achieve differential privacy is to return a query answer R that consists in the query result computed on the original data plus Laplace-distributed noise. The smaller the value of ϵ and the more sensitive the query (i.e. the larger the potential change of the query result due to the change of a single record), the greater the amount of noise required.

Dwork (2011) seems to suggest that the $O(\sqrt{n})$ accuracy provided by randomized response (Warner 1965) can be outperformed by a differentially private procedure, when she writes:

> Suppose n respondents each employ randomized response independently, but using coins of known, fixed, bias. Then, given the randomized data, by the properties of the binomial distribution the analyst can approximate the true answer to the question "How many respondents have value b?" to within an expected error on the order of $O(\sqrt{n})$. As we will see, it is possible to do much better–obtaining constant expected error, independent of n.

Yet, achieving constant error independent of n clashes with the requirement of Dinur and Nissim (2003) according to which, to prevent a (polynomial) adversary from being able to reconstruct a database based on query outputs, noise at least $O(\sqrt{n})$ is needed for a non-negligible proportion of queries. As Dwork acknowledges above, achieving $O(\sqrt{n})$ noise is precisely what randomized response does.

Hence, if a differentially private procedure offers constant error independent of n, it cannot protect against reconstruction according to DN. In fact, if a DN-adversary is allowed to submit a polynomial number of queries, say $m = O(n)$, sequential composition applies, because in general queries may be on overlapping sets of individuals. Thus, the total privacy budget ϵ must be split into chunks of ϵ/m per query. Hence, since the noise applied to each query answer is inversely proportional to its privacy budget, for the Laplace mechanism the standard deviation of the noise is directly proportional to m/ϵ and therefore $O(n)$. To summarize, if ϵ-DP is correctly applied, it protects against reconstruction because it uses $O(n)$ noise, considerably more noise than randomized response. Therefore, ϵ-DP mechanisms are likely to over-protect outputs of increasing complexity, as noted in Bach (2022).

Furthermore, when comparing RR and DP it must be noted that, even though RR can satisfy the DP requirements (Dwork 2011), RR was proposed in 1965, more than a decade before the birth of the SDC discipline and four decades before DP. In fact, RR has other properties beyond DP, such as allowing an estimation of the original distribution based on the randomized distribution.

4. The Relevace of Geography and Policy Decisions for Reconstruction

U.S. law requires that the U.S. Census Bureau not “make any publication whereby the data furnished by any particular establishment or individual ... can be identified.” (U.S. Census Bureau 2021a). Any individual with unique characteristics at the lowest geographic level at which the tables are released is at risk of reidentification; that is, any cell count of 1 is exposed to reidentification.

One of the key reasons for implementing DP in the Census context is the claim that the swapping approach used to protect previous decennial censuses was ineffective against reconstruction and reidentification. Garfinkel et al. (2019) provide a simple hypothetical example using primary and secondary suppression to highlight the danger of reconstruction. It has been shown (Muralidhar and Domingo-Ferrer 2021, 2022) that even this reconstruction would have been infeasible if primary and secondary suppression had been applied in the correct way (e.g., as described in Census methodology documentation (Dupre 2020) and related SDC literature (Antal et al. 2017; UNECE 2015)). It remains however true that publishing statistics at a detailed geographic level may facilitate reconstruction. We examine this issue in what follows.

4.1. The Impact of Geography

If statistics are released in small geographies, reconstruction can be performed using simple arithmetic. In fact, no matter whether swapping or DP is used as an SDC approach to protect tables, if total counts are exactly preserved at a small geographic level, then reconstruction is feasible (see Abowd and Hawes 2022, 8). The ease of reconstruction greatly depends on how small are the geographic areas for which exact counts are reported. More precisely, in the comparison between swapping and DP on the Census 2010 data conducted by the U.S. Census Bureau:

- When implementing swapping on the 2010 Census, total population and voting age counts were held invariant (exactly reported) at the block level (Abowd 2021a, 12).

- In contrast, when implementing differential privacy on the 2010 Census, only the state-level population was held invariant. Note that in 2017, block-level exact counts had been promised: "By agreement with the Department of Justice, the Census Bureau will provide exact counts at the Census block level..." (Dajani et al. 2017).

Now, there are over six million blocks versus only 50 states plus the District of Columbia. Thus, swapping was far more constrained than DP and, as a result, more disclosive. Eliminating block-level constraints (preserving total population and voting-age population counts) for swapping might put the privacy protection afforded by swapping on the same footing as DP). A fair comparison between swapping and DP would require the U.S. Census Bureau to report the results of their reconstruction attacks applied to both swapped data and DP-protected data from the 2010 Decennial Census when exact counts are preserved at the same geographic level. This would allow comparing the protection and the utility provided by both approaches; in particular, it would be interesting to see the extent to which reconstruction on DP-protected data can be performed in the same way described for swapping (Garfinkel et al. 2019, 34).

Although the U.S. Census Bureau claims to have performed a comparative analysis of DP against swapping and suppression, no specific comparative results are available. Only the following statement is provided: "to achieve the necessary level of privacy protection, both enhanced data swapping and suppression had severely deleterious effects on data quality and availability" (Abowd 2021a, 25).

Another concern is that even relaxing from exact count preservation to consistent count preservation at several geographic levels is problematic under DP. According to Garfinkel (2019, 59), noise can be added in all geographic levels of the Census 2020 as long as consistency is maintained. To ensure this consistency, the DAS methodology developed by the Census based on DP involves several postprocessing steps (Kenny et al. 2021).

4.2. *The Impact of the Privacy Budget on DP*

If DP is advocated as a replacement of traditional SDC methods, the privacy budget ϵ should be specified, as enjoined in Dwork et al. (2019). Taking a very small ϵ entails unaffordable utility loss, but taking ϵ very large entails very little noise addition and offers little to no protection against reconstruction, let alone reidentification (Domingo-Ferrer et al. 2021).

In fact, recent U.S. Census documents mention ϵ values as high as 19.61 in 2021 (U.S. Census 2021b) and 39.907 in 2022 (U.S. Census 2022). Let us take the 2021 ϵ value to illustrate how little privacy it achieves (the 2022 value still achieves less). We offer two different views that lead to similar conclusions:

1. First, we use the connection between DP and randomized response (Wang et al. 2016). Consider RR for a binary attribute, so that the reported randomized answer is equivalent to the true answer with probability $p \geq 0.5$ and different with probability $1-p$. Then, for any ϵ, the disclosure risk incurred by ϵ-DP is the same incurred by RR when $p = \exp(\epsilon)/(1 + \exp(\epsilon))$. Specifically, $\epsilon = 19.61$ translates to binary RR with $p = 0.99999999696$, that is, to RR reporting the original value with probability practically 1, which basically amounts to no disclosure protection being offered.

2. Alternatively, we take the Dinur and Nissim perspective. For illustrative purposes, assume that the noise added is sampled from a Laplace distribution. For $\epsilon = 19.61$, the noise is bounded in the range $[-1, 1]$ with probability higher than 0.999999997, and it is bounded in the range $[-0.5, 0.5]$ with probability higher than 0.999944825. Adding this level of noise would violate the DN requirement that the noise should be $O(\sqrt{n})$ and would allow accurate reconstruction of the data.

Worse yet, even with high values of ϵ, the utility of the released DP data can be very low in some cases, as noted in Van Riper et al. (2020), Ruggles and Van Riper (2022) and Kenny et al. (2021). This is due to: (A) sequential composition, which requires splitting the privacy budget among all released outputs that are not independent of each other (for instance, among all the cells of a table, among different geographic levels, among queries related to each other, etc.); and (B) post-processing with the Census's TopDown Algorithm (TDA), which is required in order to publish data that are consistent, integral and non-negative.

The real protection and the utility loss of the high values of ϵ being proposed should be compared to those achievable using traditional SDC methods (e.g., those employed in the 2010 Decennial Census) under the same invariance constraints.

4.3. Transparency

One of the key claims when using differential privacy is that "In turn, this allows an agency like the Census Bureau to quantify the precise amount of statistical noise required to protect privacy. This precision allows the Census to calibrate and allocate precise amounts of statistical noise in a way that protects privacy while maintaining the overall statistical validity of the data" (Abowd 2021a, 22).

In fact, this is true of any methodology, including swapping. It is possible to select the swapping parameters to: (1) include (more or less) records to be swapped, (2) the attributes to be swapped, and (3) whether the swapping is performed independently for each attribute.

Releasing the ϵ parameter used in DP, as the U.S. Census Bureau does, is certainly a step in the good direction. However, this alone does not make the protection methodology transparent. The postprocessing employed remains opaque to the users. One of the key criticisms against the swapping methodology employed until the Census 2010 was that the swapping parameter (the proportion of swapped records) was not released to the public. But the U.S. Census Bureau did release an upper bound for the proportion of swapped records. Given the simplicity of swapping, this made the procedure pretty transparent. In addition, swapping also assured that certain counts were preserved even at the block level, which afforded still greater transparency.

In our opinion, transparency is not just a matter of parameter release; it also has to do with the complexity of the approach. The more complex it is, the less transparent it is to the users. In this light, the current DP-based approach can be construed as being less transparent than simple swapping.

5. Reconstruction and Reidentification are Different

Reconstruction and reidentification are two different notions:

1. Reconstruction is only the first step in the disclosure process. Note that reconstructing from the outputs of statistical queries or from released tabulations yields reconstructed data that include no identifiers.
2. Reidentification is a second and necessary step to complete a disclosure attack. In this step, the reconstructed data are *linked* to a particular individual. To this end, the attacker needs an external data source that contains identifiers plus some attributes that can be used to link with the reconstructed data. In the worst-case scenario (most favorable to the attacker), the external data source may contain the *entire* original data with identification information. For example, this worst-case scenario makes sense if the attack is conducted by the same organization that protects the data (in order to test the quality of reconstruction).

Abowd (2021a, app. B) and Garfinkel (2019) describe the reconstruction and reidentification as follows:

- The microdata records of 308,745,538 people were "reconstructed".
- Four external commercial databases of the 2010 US population were used, which reported "Name", "Address", "Age", and "Gender" of people.
- The reconstructed records were linked to the commercial databases to obtain a linked database with "Name", "Address", "Age", "Gender", "Ethnicity", and "Race". 45% of records could be linked.
- The linked database was compared to the U.S. Census Bureau confidential data. It is claimed that the attack got all attributes in 38% of the linked records, or equivalently for 17% of the U.S. population.
- Hence, the authors claim reidentification of 17% of the U.S. population, although Garfinkel (2019) concedes that an outside attacker would not know which reidentifications were correct.

After that, the authors go on to criticize as flawed the protection system used in the 2000 and 2010 Censuses, which relied on traditional SDC techniques. This is used as a justification to move to formal privacy, which amounts to DP. In Garfinkel (2019) it is explained that choosing the privacy budget ϵ is a public policy choice.

5.1. Issues with Reidentification Claims

There are several issues with the claims in Abowd (2021a, app. B):

1. It is unclear what "reconstructing" the microdata of 308,745,538 people signifies. According to Van Riper et al. (2020), it amounted to re-generating microdata records from published census block and tract tabulations, that is, from frequency tables with attributes "Census block", "Age", and "Gender"). This is not true reconstruction as DN describe. Note that, in general, the re-generation of microdata from a frequency table is not unique, because a frequency table contains less information than the microdata it was computed from. Hence, just re-generating one of the microdata sets that are compatible with a certain frequency table does not qualify as reconstruction of the original data: in DN's notion of reconstruction, the accuracy bounds are essential, and no such bounds are given for the Census so called reconstruction (Muralidhar 2022).

2. Reidentification means being able to link the records in anonymized microdata with the corresponding records in an external data set containing identifiers and covering a similar population. This is not reconstruction. Since the attack was based on microdata re-generated from frequency tabulations, proper reidentification could only be conducted from those cells with count 1. In all other cases, unequivocal reidentification is impossible.
3. It has been known at least since Sweeney (2000) that matching a database containing demographic attributes such as “Municipality of residence”, “Birthdate”, and “Gender” against an external database containing those same attributes plus identifiers for the same population is likely to yield a high proportion of reidentifications. In fact, Ruggles and Van Riper (2022) show using a simulation that most matches reported by the U.S. Census Bureau experiment at a block level would be expected randomly and *thus fail to demonstrate a credible threat to confidentiality*. Hence, the use of DP may not be necessary. Even if the threat to confidentiality was credible, it is unclear that the Census’s new DP-based TDA algorithm offers the best protection. In Francis (2022) it is shown that race and ethnicity can be inferred with more precision and less prior knowledge from TDA outputs than from the outputs of the Census previous protection algorithm.
   An alternative has also been proposed by Ruggles (2021) to investigate the impact of reconstruction on reidentification. The idea is to first match the external database to the reconstructed census data. That yields a certain matching rate r. Then take those unmatched records from the external database and compare them by block ID and PIK (the Protected Identification Key created by the Census Bureau for each original record) to the Census Edited File (the original confidential data). Let r' be the reidentification rate resulting from this comparison. If $r \approx r'$, then database reconstruction has little or no impact on reidentification; to demonstrate that reconstruction increases the reidentification risk, r should be substantially greater than r'. The U.S. Census Bureau is yet to make this comparison.

The above issues clearly show that, rather than focusing on reidentification, the Census experiment focuses on finding (non-unique) candidate reconstructions. We show next that (mis)interpreting reconstruction as reidentification may in some situations overstate and in other situations understate the real risk of reidentification.

5.2. *Misinterpreting Reconstruction as Reidentification May Overstate or Understate the Reidentification Risk*

Recall that in the U.S. Census Bureau’s “reidentification” procedure described in Abowd (2021a) and Garfinkel (2019), and summarized at the beginning of Section 5, reconstructed microdata reporting,“Gender”, “Age”, “Race”, and “Ethnicity” are linked to an external commercial database reporting “Name”, “Address”, “Age”, and “Gener”. Thus, linkage is performed using the “Age” and “Gender” attributes. As a result of linkage, a linked database is obtained that reports “Name”, “Address”, “Age”,“Gender”, “Race”, and “Ethnicity”.

Consider three scenarios at the block level:

- **Scenario 1.** Block whose reconstructed data consist of ten individuals with "Age" = 44, "Gender" = Male, "Race" = White and "Ethniciy" = Not_Hispanic. The commercial database contains "Name", "Address", "Age" = 44 and "Gender" = Male for all individuals in this block.
- **Scenario 2.** Same data as in Scenario 1, but with an additional attibute *Relationship*, which according to Garfinkel (2019) is also collected for each person in a block and can take 17 different values. Assume that in, this scenario, each of the ten persons in the block has a different "Relationship" value. Like in Scenario 1, the commercial database contains "Name", "Address", "Age" = 44 and "Gender" = Male for all individuals in this block.
- **Scenario 3.** Block whose reconstructed data consist of ten individuals with "Age" = 44, "Gender" = "Male", such that such that all ten of these individuals belong to different ("Race", "Ethnicity") combinations. The commercial database contains "Name", "Address", "Race" and "Ethnicity" for all individuals in this block, but no "Age" or "Gender".

In Scenario 1, the U.S. Census Bureau's procedure would yield a 100% reconstruction, because the attacker would always be able to associate the correct "Race" and "Ethnicity" to the ten names and addresses in the block. Yet, claiming that this 100% reconstruction amounts to 100% reidentification is patently incorrect, because the attacker has no way to confirm the identification of ("Name", "Address") for the ten individuals who are indistinguishable from one another – "reidentification" in this case can be attributed to the homogeneity of a block and is not a true reidentification. The U.S. Census Bureau document McKenna and Haubach (2019) states that "it is necessary to verify the proposed matches by comparing the suppressed identities in the microdata with the identities in the external data set to see if the matches are true matches or false matches."

The above point that correctly reconstructing "Ethnicity" and "Race" does not amount to reidentification becomes apparent in Scenario 2. When the attribute "Relationship" is added with different values for all ten individuals, it becomes clear that the reidentification probability for any specific individual is in fact 1/10.

In Scenario 3, both the probability of correct reconstruction and the probability of correct reidentification are 1, but for different reasons:

- Since all individuals have the same combination of ("Age", "Gender"), reconstructing the values of these attributes for the ten individuals is trivial, which yields 100% reconstruction. Note that if not all individuals had the same combination, then the probability of correct reconstruction would be less than 1.
- Since all individuals have different combinations of ("Race", "Ethnicity"), unequivocally linking each of the ten records in the reconstructed data to its corresponding record in the commercial database is straightforward, which yields 100% reidentification.

The above shows that reconstruction and reidentification are different notions. The bottom line is as follows: whereas *reconstruction is helped by homogeneity of the missing confidential attributes, reidentification is helped by heterogeneity of the quasi-identifiers through which linkage is performed.*

Hence, the reconstruction procedure described in Abowd (2021a) and Garfinkel (2019) does not yield an appropriate measure of reidentification risk. In fact, it is likely to overstate the reidentification risk (as in Scenario 1), since at the block level ("Race", "Ethnicity") can be expected to be fairly homogeneous, which makes Scenario 1 more likely than Scenario 3 (Ruggles and Van Riper 2022).

Interestingly, researchers at the U.S. Census Bureau have performed in the past extensive research in reidentification risk (e.g., Winkler 1999). To assess the true risk of reidentification, it is necessary to assume the following. At the block level, the reconstructed data consist of ("Block ID", "Gender", "Age", "Race", "Ethnicity", and "Relationship".) and the attacker has the attributes ("Name", "Address", "Block ID", "Gender", "Age", "Race", "Ethnicity", and "Relationship"). The objective of reidentification is to uniquely link a record from the (unidentified) reconstructed data to a record in the (identified) attacker's data thereby attaching ("Name", "Address") to the reconstructed data. *Such a procedure will correctly assess the reidentification risk in the scenarios described above*; as mentioned above and in McKenna and Haubach (2019), once the linkage is established, reidentification needs to be validated by checking that the linkage is unique and that identities (name and address in this case) match between the attacker's record and the original record to which the unidentified reconstructed record corresponds. Reconstructing unidentified records, in itself, does not pose a real disclosure threat. *Reconstruction* in the DN sense also requires to be supplemented by correct re-identification (Bach 2022). Only then does it constitute real disclosure.

6. Conclusions and Future Work

In this article, we have reassessed the feasibility of reconstructing a data set based on the outputs of statistical queries computed on it. The danger of reconstruction has been cited as an argument to justify the use of differential privacy in official statistics, most notably in the case of the 2020 Census of the U.S.A. Using DP, however, will most likely result in a decrease of the utility of the statistical outputs of that Census. This article has investigated to what extent reconstruction is a real danger.

We first examined the state of the art in reconstruction theory – Dinur and Nissim's framework – and we concluded that local or input protection appear as good ways to resist reconstruction. If the U.S. Census Bureau were to stick to the so-called CD-model and produced locally protected or input-protected data (e.g., using RR or microdata masking discussed in Hundepool et al. (2012), or the methods used in the 2010 Census), then reconstruction would not be a real danger: at most the attacker would be able to reconstruct the locally protected or the input-protected data, rather than the original data. Differential privacy is also an option, but it may add more noise than strictly required to counter reconstruction, thereby leading to unnecessary utility loss (Dove 2021; Hotz et al. 2022; Bach 2022), or it may offer less protection than previous approaches (Francis 2022).

We then highlighted the relevance for protection of the geographic levels at which exact population counts are preserved. No matter the SDC methods used, preserving counts in small geographies facilitates reconstruction, while not preserving counts in small geographies goes a long way towards avoiding reconstruction, but also reducing utility.

Finally, we have warned against using the amount of reconstruction as a measure of reidentification risk, which results in exaggerated reidentification risk. Whereas reconstruction requires only query outputs or tabular unidentified outputs and is favored by the homogeneity of the missing attribute values, reidentification also needs external identified sources and is favored by the heterogeneity of the values of the attributes used for linking with those sources.

An additional concern are the successive increases of the value of ϵ during the process. Increasing ϵ implies a loss of privacy. The U.S. Census Bureau started with Laplace noise addition with $\epsilon = 4.5$ and subsequently increased to $\epsilon = 10.2$. To further improve the utility of the released data, the Bureau adopted zero-concentrated DP (with noise from a discrete Gaussian distribution, (Bun and Steinke 2016)) in place of Laplace noise. The parameter ρ of zero-concentrated DP can be used to compute equivalent values ϵ and δ for (ϵ, δ)-DP. For a given ρ there are many equivalent combinations (ϵ, δ). However, for fixed δ (in the case of the Census it is $\delta = 10^{-10}$), then each ρ has a single equivalent ϵ. In 2021, this equivalent global ϵ was 19.61 (U.S. Census 2021). This value was further revised to a global $\epsilon = 39.907$ in year 2022 (U.S. Census 2022). This has a great impact on privacy. The privacy level associated with $\epsilon = 39.907$ is worse than the privacy level given by $\epsilon = 4.5$ by a factor $e^{39.907}/e^{4.5} = 2.382 \times 10^{15}$. Referring to Apple's use of $\epsilon = 14$, Frank McSherry, one of the inventors of DP, commented that it was "pointless" in terms of privacy (Greenberg 2017).

Actually, $\epsilon = 39.907$ is over 1.78×10^{11} times worse than $\epsilon = 14$. In Abowd (2021b), the then Census's Chief Scientist said about ϵ that "specifically it limits the statistical power of all possible tests for whether a particular individual's data record (or portions thereof) was used to produce a collection of statistics versus the record of another, arbitrary individual." With $\epsilon = 39.907$ used in 2022, this implies that the probability that a particular individual's data record was used can be over 2.14×10^{17} times higher versus the record of another, arbitrary individual. Since the current U.S. population is only 3.31×10^8, with an $\epsilon = 39.907$ any target U.S. inhabitant might be reidentifiable. In fact, this also held for the $\epsilon = 19.61$ used in 2021.

Even with this relaxation of the value of ϵ, there are still very serious utility concerns. Consider the following report in the New York Times (Wines 2022): "According to the 2020 census, 14 people live there (in Census Block 1002 in downtown Chicago) −13 adults and one child. Also according to the 2020 census, they live underwater. Because the block consists entirely of a 700-foot bend in the Chicago River." Or this analysis from Cornell University (Cornell 2021) of the 2021 Census DAS release for New York state which shows that in 6.1% of the blocks, the household population is greater than 0, but the number of occupied houses is 0; in 2.5% of the blocks, the household population is less than the number of occupied houses (which means there is less than 1 person per household); and in 0.8% of the blocks, the household population is 0, but the number of occupied houses is greater than 0. These results are impossible and would not have occurred in the 2010 Census. Thus, even with large ϵ, the differentially private noise addition procedure is not capable of providing accurate and consistent results. In fact, the U.S. Census Bureau recently announced that "for the time being, the ACS PUMS (American Community Survey Public Use Microdata Sample) data product will still be protected using traditional disclosure avoidance methods", since it is "not clear that differential privacy would ultimately be the best option." (Daily 2022)

In this study, we conclude that the concern of database reconstruction resulting in mass disclosure is unwarranted. We believe that these claims are based on a comparison that is incomplete and opaque – only the Census Bureau can assess or verify the true reidentification results. Other researchers, some of them mentioned above, have raised serious concerns regarding the accuracy and consistency of the output. Hence, it is not clear that differential privacy is the best option for the 2020 decennial census data. We suggest a comprehensive, independent, fully documented, peer-reviewed assessment of the efficacy of alternative methods.

7. References

Abowd, J.M. 2017. "Research data centers, reproducible science, and confidentiality protection: the role of the 21st century statistical agency." Presentation to Summer DemSem, Wisconsin Federal Statistical RDC. June 5, Madison, Wisconsin. Available at: https://www2.census.gov/cac/sac/meetings/2017-09/role-statistical-agency.pdf (accessed February 2023).

Abowd, J.M., I.M. Schmutte, W.N. Sexton, and L. Vilhuber. 2019. "Why the economics profession cannot cede the discussion of privacy protection to computer scientists." Presentation to The Future of Economic Research under Rising Risks and Costs of Information Disclosure. Allied Social Science Associations Annual Meetings. January 5, Atlanta, Georgia. https://ecommons.cornell.edu/handle/1813/60836 (accessed February 2023).

Abowd., J.M. 2021a. *Declaration of John M. Abowd. Case no. 3:21-CV-211-RAH-ECM-KCN.* U.S. District Court for the Middle District of Alabama. Availabe at: https://censusproject.files.wordpress.com/2021/04/2021.04.13-abowd-declaration-alabama-v.-commerce-ii-final-signed.pdf (accessed June 2023).

Abowd, J.M. 2021b. *Supplemental Declaration of John M. Abowd. Case no. 3:21-CV-211-RAH-ECM-KCN.* U.S. District Court for the Middle District of Alabama. Available at: https://www.brennancenter.org/sites/default/files/2021-06/M.D.%20Ala.%2021-cv-00211%20dckt%20000116_001%20filed%202021-04-26%20Abowd%20declaration.pdf (accessed June 2023).

Abowd J.M., and M.B. Hawes. 2022. *Confidentiality Protection in the 2020 US Census Population and Housing*. Available at: https://arxiv.org/pdf/2206.03524.pdf (accessed February 2023).

Adam, N.R., and J.C. Worthmann. 1989. "Security-control methods for statistical databases: a comparative study." *ACM Computing Surveys* 21(4): 515–556. DOI: https://doi.org/10.1145/76894.76895.

Alabama. 2021. *Alabama* v. *U.S. Dep't of Commerce. 2021*. Brennan Center for Justice. https://www.brennancenter.org/our-work/court-cases/alabama-v-us-dept-commerce (accessed February 2023).

Antal, L., T. Enderle, and S. Giessing. 2017. *Statistical disclosure control methods for harmonised protection of census data*. Deliverable D3.1 Part I, Eurostat contract

"Harmonised Protection of Census Data in the ESS", Available at: https://ec.europa.eu/eurostat/cros/system/files/methods_for_protectingcensus_data.pdf (accessed February 2023).

Associated Press. 2021. "16 states back Alabama's challenge to Census privacy tool." *U.S. News*. Available at: https://www.usnews.com/news/us/articles/2021-04-13/16-states-back-alabamas-challenge-to-census-privacy-tool (accessed February 2023).

Bach. 2022. "Differential privacy and noisy confidentiality concepts for European population statistics." *Journal of Survey Statistics and Methodology*, 10: 642–687. DOI: https://doi.org/10.1093/jssam/smab044.

Bun, M., and T. Steinke. 2016. "Concentrated differential privacy: simplifications, extensions, and lower bounds." In. Theory of Cryptography Conference-TCC, October 31–November 3, Beijing, China. Springer: 635-658. DOI:https://doi.org/10.1007/978-3-662-53641-4_24. https://link.springer.com/chapter/10.1007/978-3-662-53641-4_24.

Cornell. 2021. Census 2020 results: Data and Analyses for New York from the data products as they are released over time by the U.S. Census Bureau. Cornell Program on Applied Demographics. Available at: https://pad.human.cornell.edu/census2020/index.cfm#das (accessed February 2023).

Dalenius. T. 1977. "Towards a Methodology for Statistical Disclosure Control." *Statistisk Tidskrift* 15: 429–444. Available at: https://ecommons.cornell.edu/bitstream/handle/1813/111303/dalenius-1977.pdf?sequence = 3&isAllowed = y (accessed June 2023).

Daily, D. 2022. *Disclosure avoidance protections for the American Community Survey*. U.S. Census Bureau. Available at: https://www.census.gov/newsroom/blogs/random-samplings/2022/12/disclosure-avoidance-protections-acs.html (accessed February 2023).

Dajani, A.N., A.D. Lauger, P.E. Singer, D. Kifer, J.P. Reiter, A. Machanavajjhala, S.L. Garfinkel, S.A. Dahl, M. Graham, V. Karwa, H. Kim, P. Leclerc, I.M. Schmutte, W.N. Sexton, L. Vilhuber, and J.M. Abowd. 2017. "The modernization of statistical disclosure limitation at the U.S. Census Bureau. In Census Scientific Advisory Committee Meeting, Sepember 14–15, Suitland MD, USA. Available at: https//www.census.gov/library/video/2017/2017-09-sac.html (accessed February 2023).

Denning, D.E., and J. Schlorer. 1980. "A fast procedure for finding a tracker in a statistical database." *ACM Transactions on Database Systems* 5(1): 88–102. DOI: https://doi.org/10.1145/320128.320138.

Dinur, I., and K. Nissim. 2003. "Revealing information while preserving privacy." In. Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, Jun. 9–12. San Diego CA, USA.: 202–210, DOI: 10.1145/773153.773173; https://dl.acm.org/doi/10.1145/773153.773173.

Domingo-Ferrer, J.D. Sánchez, and A. Blanco-Justicia. 2021. "The limits of differential privacy (and its misuse in data release and machine learning)." Communications of the ACM 64(7): 33–35. DOI: https://doi.org/10.1145/3433638.

Dove, I. 2021. *Applying differential privacy protection to ONS mortality data, pilot study*. Office for National Statistics. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/methodologies/applyingdifferentialprivacyprotectiontoonsmortalitydatapilotstudy (accessed June 2023).

Duncan, G.T., S.A. Keller-McNulty, and S. Lynne. 2001. *Disclosure Risk vs. Data Utility: The RU Confidentiality Map.* National Institute of Statistical Sciences. Technical Report no. 121. Available at: https://www.niss.org/sites/default/files/technicalreports/tr121.pdf (accessed February 2023).

Dupre, S. 2020. *Disclosure avoidance and the Census. Select Topics in International Censuses*. U.S. Census Bureau. Available at: https://www.census.gov/content/dam/-Census/library/working-articles/2020/demo/disclosure_avoidance_and_the_census_-brief.pdf (accessed February 2023).

Dwork, C. 2011. "A firm foundation for private data analysis." *Communications of the ACM* 54(1): 86–95. DOI: https://doi.org/10.1145/1866739.1866758.

Dwork, C.N. Kohli, and D. Mulligan. 2019. "Differential privacy in practice: expose your epsilons!" *Journal of Privacy and Confidentiality* 9(2): 9–22. DOI: https://doi.org/10.29012/ jpc.689.

Dwork, C., F. McSherry, K. Nissim, and A. Smith. 2006. "Calibrating noise to sensitivity in private data analysis." In Theory of Cryptography Conference – TCC 2006, March. 4–7, New York NY USA. Springer: 265–284. DOI: https://doi.org/10.1007/11681878 https://link.springer.com/chapter/10.1007/11681878_14.

Dwork, C. and A. Roth. 2014. *The Algorithmic Foundations of Differential Privacy*. Now Publishers. DOI: https://doi.org/10.1561/0400000042.

Francis, p. 2022. "A note on the misinterpretation of the US Census re-identification attack." In *Privacy in Statistical Database – PSD 2022*, Sepember 21–23, Paris, France. Springer: 299–311. DOI: https://doi.org/101007/9783-031-13945-1 https://link.springer.com/chapter/10.1007/978-3-031-13945-1_21.

Garfinkel, S. 2019. "Deploying differential privacy for the 2020 Census of Population and Housing." In Privacy Enhancing Technologies Symposium–PETS 2019, July 16–20, Stockholm, Sweden. Available at: https://simson.net/page/Main_Page (accessed February 2023).

Garfinkel, S., J.M. Abowd, and C. Martindale. 2019. "Understanding database reconstruction attacks on public data." *Communications of the ACM* 62(3): 46–53. DOI: https://doi.org/10.1145/3291276.3295691.

GDPR. 2016. *General Protection Regulation.* Regulation (EU) 2016/679. Available at: https://gdpr-info.eu (accessed February 2023).

Greenberg, A. 2017. "How one of Apple's key privacy safeguards falls short." *Wired.* Available at: https://www.wired.com/story/apple-differential-privacy-shortcomings/ (accessed February 2023).

Hotz, V.J., C.R. Bollinger, T. Komarova, C.F. Manski, R.A. Moffitt, D.Nekipelov, A. Sojourner, and B.D. Spencer. 2022. "Balancing data privacy and usability in the federal statistical system." *PNAS* 119(31): e2104906119. DOI: https://doi.org/10.1073/pnas.21049 06119.

Hundepool, A.J. Domingo-Ferrer, L. Franconi, K. Spicer, P.-P. De Wolf, S. Giessing, and E. Schulte Nordholt. 2012. *Statistical Disclosure Control.* Wiley. DOI: https://doi.org/10.1002/9781118348239.

Kenny, C.T.S. Kuriwaki, C. McCartan, E.T.R. Rosenman, T. Simko, and K. Imai. 2021. "The use of differential privacy for census data and its impact on redistricting: the case of the 2020 U.S. Census." Science Advances 7(41): eabk3283. DOI: https://doi.org/10.1126/sciadv.abk3283.

McKenna, L., and M. Haubach. 2019. *Legacy techniques and current research in disclosure avoidance at the U.S. Census Bureau.* Research and Methodology Directorate, U.S. Census Bureau. Available at: https://www.census.gov/library/working-articles/2019/adrm/CED-WP-2019-005.html (accessed February 2023).

Muralidhar, K. 2022. "A re-examination of the Census Bureau reconstruction and reidentification attack. In Privacy in Statistical Database – PSD 2022, September 21–23, Paris, France. Springer: DOI: https://doi.org/10.1007/978-3-031-13945-1 https://link.springer.com/chapter/10.1007/978-3-031-13945-1_22.

Muralidhar, K., and J. Domingo-Ferrer. 2021. "Database reconstruction is very difficult in practice!" In 2021 Joint UNECE/Eurostat Expert Meeting on Statistical Data Confidentiality, december 1–3, Poznan, Poland. Available at: https://unece.org/sites/default/files/2021-12/SDC2021_Day1_Muralidhar_AD.pdf. (accessed February 2023).

Muralidhar, K., and J. Domingo-Ferrer. 2022. "Census reconsiderations". *Communications of the ACM* 65(6): 11. DOI: https://doi.org/10.1145/3532630.

Muralidhar, K., and R. Sarathy. 2009. "Privacy violations in accountability data released to the public by state educational agencies." In Federal Committee on Statistical Methodology Research Conference, November 2–4, Washington D.C. USA. Available at: https://www.researchgate.net/profile/Rathindra-Sarathy/publication/273448878_Privacy_Violations_in_Accountability_Data_Released_to_the_Public_by_State_Educational_Agencies_Rathindra_Sarathy/links/5501d43e0cf231de076ca7b3/Privacy-Violations-in-Accountability-Data-Released-to-the-Public-by-State-Educational-Agencies-Rathindra-Sarathy.pdf (accessed June 2023).

Percival, K. 2021. *Court rejects Alabama challenge to Census plans for redistricting and privacy*. Brennan Center for Justice. Available at: https://www.brennancenter.org/our-work/analysis-opinion/court-rejects-alabama-challenge-census-plans-redistricting-and-privacy (accessed February 2023).

Ruggles. S. 2021. *Personal communication, November 9.*

Ruggles, S., and D. van Riper 2022. "The role of chance in the Census Bureau database reconstruction experiment." *Population Research and Policy Review* 41: 781–788. DOI: https://doi.org/10.1007/s11113-021-09674-3.

Schneider, M. 2022. "Researchers ask Census to stop controversial privacy method." AP News., Available at: https://apnews.com/article/census-2020-us-bureau-government-and-politics-20e683c71eeb62ee4b7792d7d8530419 (accessed February 2023).

Sweeney, L. 2000. *Simple demographics often identify people uniquely.* Carnegie Mellon University, Data Privacy Working Paper 3. Available at: https://ggs685.pbworks.com/wZfile/fetch/94376315/Latanya.pdf (accessed February 2023).

Traub, J.F., Y. Yemini, and H. Wozniaknowski. 1984. "The statistical security of a statistical database." *ACM Transactions on Database Systems*, 9(4): 672–679. DOI: https://doi.org/10.1145/1994.383392.

UNECE-CES. 2015. (United Nations Economic Commission for Europe – Conference of European Statisticians) *Recommendations for the 2020 Censuses of Population and Housing*. United Nations Publications, New York, NY. Available at: https://unece.org/DAM/stats/publications/2015/ECECES41_EN.pdf (accessed February 2023).

U.S. Census Bureau. 2021a. *Legal Authority and Policies for Data Linkage at Census*. Available at: https://www.census.gov/about/adrm/linkage/about/authority.html (accessed June 2023).

U.S. Census Bureau. 2021b. *Census Bureau sets key parameters to protect privacy in 2020 Census results.* Release Number CB21-CN.42. Available at: https://www.census.gov/-newsroom/press-releases/2021/2020-census-key-parameters.html (accessed February 2023).

U.S. Census Bureau. 2022. *Privacy loss Budget Allocation.* Available at: https://www2.census.gov/programs-surveys/decennial/2020/program-management/-data-product-planning/2010-demonstration-data-products/02-Demographic_and_Housing_Characteristics/2022-03-16_Summary_File/2022-03-16_Privacy-Loss_Budge-t_Allocations.pdf (accessed February 2023).

Van Ripper, D., T. Kugler, and S. Ruggles. 2020. “Disclosure avoidance in the Census Bureau’s 2010 demonstration data product.” In Privacy in Statistical Databases – PSD 2020, September 23–25, 2020, Tarragona, Catalonia. Springer: 353–368. DOI: https://doi.org/10.1007/978-3-030-57521-2 https://link.springer.com/chapter/10.1007/978-3-030-57521-2_25.

Y. Wang, Y., X. Wu, and D. Hu. 2016. “Usings randomized response for differential privacy preserving data collection.” In Proceedings of the EDBT/ICDT 2016 Joint Conference, March 15–18, Bordeaux, France. DOI: https://doi.org/10.5441/002/edbt.2016.01; https://ceur-ws. org/Vol- 1558/article35.Pdf.

Warner, S.L. 1965. “Randomized response: a survey technique for eliminating evasive answer bias.” *Journal of the American Statistical Association*, 60(309): 63–69. DOI: https://doi.org/10.1080/01621459.1965.10480775.

Wines, M. 2022. “The 2020 Census suggests that people live underwater.” There’s a reason. *The New York Times*, April 21. https://www.nytimes.com/2022/04/21/us/census-data-privacy-concerns.html (accessed February 2023).

Winkler, W. 1999. *The state of record linkage and current research problems.* Technical report, Statistical Research Division, U.S. Census Bureau. Available at: https://courses.cs.Washington.edu/courses/cse590q/04au/articles/Winkler99.pdf (accessed February 27, 2023).

Received February 2022
Revised August 2022
Accepted March 2023

Journal of Official Statistics, Vol. 39, No. 3, 2023, pp. 399–410, http://dx.doi.org/10.2478/JOS-2023-0018

Comment to Mulalidhar and Domingo-Ferrer (2023) – Legacy Statistical Disclosure Limitation Techniques Were Not An Option for the 2020 US Census of Population And Housing

Simson Garfinkel[1]

The Article Database Reconstruction is Not So Easy and Is Different from Reidentification, by Krish Muralidhar and Josep Domingo-Ferrer, is an extended attack on the decision of the U.S. Census Bureau to turn its back on legacy statistical disclosure limitation techniques and instead use a bespoke algorithm based on differential privacy to protect the published data products of the Census Bureau's 2020 Census of Population and Housing (henceforth referred to as the 2020 Census). This response explains why differential privacy was the only realistic choice for protecting sensitive data collected for the 2020 Census. However, differential privacy has a social cost: it requires that practitioners admit that there is inherently a trade-off between the utility of published official statistics and the privacy loss of those whose data are collected under a pledge of confidentiality.

Key words: Differential privacy; 2020 U.S. Census; Statistical disclosure limitation; Statistical disclosure avoidance; topdown algorithm.

Reconstructing microdata for the US 2010 Census of Population and Housing that is consistent with the published statistical tables is not easy, but it is straightforward and within the capabilities of a welltrained and resourced data scientist. I know this; I did it multiple times when I was part of the team that performed the reconstruction of the 2010 Census data at the U.S. Census Bureau. Tamara Adams performed the original reconstruction, developing software that analyzed the published statistics and created a series of mixed-integer linear programs that a commercial optimizer then solved to create microdata consistent with the published constraints (Leclerc 2019). My primary contribution was to improve the system's efficiency and scalability. When we were done, it was possible to reconstruct all 308,745,538 person-level records in a few hours by typing a single command.

Others have followed our lead. "Following the details available in public presentations from Mr. Abowd and his colleagues, we were able to perform our own reconstruction experiment on Manhattan," Hansen wrote in The New York Times. "We used simple tools like R and the Gurobi Optimizer; and within a week we had our first results." (Hansen 2018) Another reconstruction can be found in Dick et al. (2023).

Keyes and Flaxman (2022) also reconstructed the 2010 data for each census block of Texas, but with a twist: they also reconstructed two sets of 2020 data for each block, and then looked for families in the 2010 data with pre-teens that had teenagers in the 2020 data of a different sex. "Census Data Put Trans Children at Risk," was their conclusion. Using

[1] 480 Main Street, New York, NY 10044. U.S.A. Email: simsong@acm.org

Acknowledgments: The author would like to thank the anonymous reviewers for their constructive comments.

synthetic 2020 data based on the 2010 Census swapping mechanism, 605 trans kids appeared in the synthetic data; had these data been real, hundreds of families would have be at risk, since the Texas Department of Family and Protective Services has been ordered by Texas Governor Greg Abbot to investigate families that provide gender-afirming care for trans youth (Chappell 2020). The threat decreased significantly using a reconstruction based on the actual 2020 data that the Census Bureau published.

Database reconstruct is different from reidentification, as Mulalidhar and Domingo-Ferrer (2023) makes abundantly clear. However, as exemplified by Texas, there are many ways that even unidentified census microdata with block-level geography can cause harm. This is why such data are protected by US law (Abowd et al. 2020). When the 2010 data tables were published, Census Bureau leadership did not think reconstruction possible. When faster computers and improved algorithms proved otherwise, it had to radically improve the statistical disclosure limitation (SDL) approaches used for the decennial census.

Reidentified microdata can be even more harmful. The Census Bureau's reidentification experiment showed that it is relatively straightforward to take the 2010 reconstructed data and match it up with commercial data that was purchased in 2010 (Hawes 2021a). The commercial data includes name, address, age and sex. The reconstructed data includes block, age, sex and race. These records can be linked, making it possible to annotate the commercial names and street addresses with race. In the US there is a long history of mixed-race families being targeted for violence. Even if "more than half of the matches are incorrect, and an external attacker has no means of confirming them" (Jarmin 2019), a group of hoodlums out to terrorize mixed race families can easily confirm the putative matches by visiting the addresses with the intent to commit acts of violence.

Mulalidhar and Domingo-Ferrer argue that the Census Bureau should have used legacy SDL techniques instead of a bespoke algorithm based on difierential privacy (DP) (Dwork et al. 2006) to protect the published data products of the 2020 Census of Population and Housing. This response provides additional context that calls into question Mulalidhar and Domingo-Ferrer's argument, and suggests a way forward for those of us who seek to improve the science of privacy protection in the production of official statistics.

NOTE: The views expressed in this article are those of the author, and do not represent the policy of the Census Bureau, the Department of Commerce, or the US Government. Also, as the author is no longer employed by the Census Bureau, this response does not present information about the Census Bureau's decision making process that has not already been placed into the public domain.

1. A Brief Introduction to Differential Privacy

Dwork et al. (2006) introduced DP to formally recognize the amount of privacy loss that invariably results when a statistics agency makes public statistics based on confidential data. DP also provides an accounting system for tracking cumulative and per-statistic privacy loss. DP is not a specific method for privacy protection. Instead, it is a mathematical definition that privacy protection *mechanisms* can satisfy. The following definition of what is now called pure (ϵ) and approximate (ϵ, δ) DP appears in Dwork and Roth (2014):

1.1. Definition 2.4 (Differential Privacy)

A randomized algorithm $\mathcal{M}$ with domain $\mathbb{N}^{|X|}$ is (ϵ, δ)-difierentially private [approximate DP] if for all $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$ and for all $x, y \in \mathbb{N}^{|X|}$ such that $||x-y||_1 \leq 1$:

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\epsilon)\Pr[\mathcal{M}(y) \in \mathcal{S}] + \delta \tag{1}$$

where the probability space is over the coin flips of the mechanism $\mathcal{M}$. If $\delta = 0$, we say that $\mathcal{M}$ is ϵ-differentially private [pure DP].

The definition considers how the output of a mechanism $\mathcal{M}$ is different for *any* two datasets, x and y, that differ in the value of a single element. This is the fundamental intuition on which DP is based: publishing an analysis of a dataset containing confidential data cannot impact a person's privacy if their data are not in the dataset. This is the source of the word "differential."

Observe that $\epsilon = 0 \Rightarrow \exp(\epsilon) = 1$, which means that when $\epsilon = 0$, mechanism $\mathcal{M}$ produces the same output for datasets x and y, implying that $\mathcal{M}$ ignores each person's data in the dataset. Thus, data subjects in x have no privacy loss when $\epsilon = 0$. Likewise, when $\epsilon = \infty$, the probability distribution of $\mathcal{M}(x)$ and $\mathcal{M}(y)$ may be arbitrarily different, which means that any person's data might be revealed. Thus, as $\epsilon \longrightarrow \infty$, there is a higher chance that an attacker will learn more aspects of any individual's data.

It turns out that randomized response (RR), which was known for decades before the introduction of DP, satises the DP definition. This allows us to use the tools of DP to easily analyze arbitrarily complex applications of RR and gain new insights into RR's potential for privacy protection.

Two specific mechanisms that implement DP are widely discussed in the literature: the "Laplace mechanism" (which uses noise from the double exponential distribution) implements ϵ-DP; and the "Gaussian mechanism" (which uses Gaussian noise) implements (ϵ, δ)-DP. Both add statistically calibrated noise to the output of an arbitrary query function f to make that function differentially private. The noise is calibrated using ϵ and a second parameter known as the "sensitivity", typically denoted as Δf. The sensitivity is the range of any individual's data on function f. For a counting query (e.g. "How many people live on this block?"), $\Delta f = 1$. Computing the sensitivity for other kinds of queries can be substantially more complex (Dwork and Roth 2014).

As Mulalidhar and Domingo-Ferrer make clear, the problem area addressed by DP is the same problem area addressed by traditional SDL techniques. But traditional SDL lacks formal definition and hence only provides heuristic guarantees which are not ground in theory, while DP's definition provides rigorous semantic guarantees regarding the protection of individual-specific information (Shlomo 2018).

2. The Advantages of Differential Privacy

DP addresses two limitations of traditional SDL: the lack of composability and the inability to make finegrained decisions regarding the social benefit of a specific statistical release and cost to individuals whose data were used.

2.1. Composability

Traditional SDL techniques such as suppression don't measure privacy loss. Instead, they have an objective: that confidential data cannot be retrieved using the published data in the current or any future data release.

With traditional SDL, each release made from an underlying confidential dataset must be coordinated; without coordination, a future release might inadvertently undo the protections of a data set already in the public domain.

Denning and Schlörer (1980) coined the term "tracker" for a data intruder that monitors the output of a statistical database to violate the stated privacy goals. The only way to protect a statistical database against a tracker is through "query auditing". Kleinberg et al. (2000) proved that query auditing is NP-hard, meaning the work effort required to properly audit appears to increase exponentially with each additional query that must be considered. Even worse, auditing itself can be a major source of privacy leakage, since the refusal to answer a query inherently releases information (Kenthapadi et al. 2013).

Official statistics agencies have recognized that the lack of composability causes a significant problem when attempting to extend traditional cell suppression techniques to large scale efforts.

There are three problems with suppression that precluded its use for the 2020 Census. First, the program that solves for the suppression pattern is not guaranteed to terminate. Second, rigorously applied primary and secondary suppression does not guarantee privacy indefinitely, as the suppressed values can be learned by other means at some point in the future. And if a future data release leaks a suppressed value, it potentially reveals many other values as well, especially if efforts were made to find and implement optimal suppression patterns. (Proof: A suppression pattern is optimal if it suppresses the minimum number of cells to achieve a specific privacy goal. If a suppressed value is later revealed, it is as if that value was not suppressed. As such, the once-optimal pattern is now suppressing one less than the minimum number of cells required to achieve the privacy goal. The goal is thus not achieved, and information that was supposed to be kept confidential is now revealed.)

The third problem with suppression is that data users had voiced strong preference for a confidentiality protection system that permitted publication of all cells, particularly in the data used for redistricting (McKenna 2018). The inability to query audit the 2010 Census tables before publication was precisely the problem demonstrated by the reconstruction: by combining data from multiple tables in a reconstruction attack, we–and others–learned information that the Census Bureau had previously pledged to keep confidential (Hawes 2021b).

2.2. Tunability

The second advantage is that DP is tunable: the amount of privacy protection for any given release can be tuned from no privacy protection (when $\epsilon = \infty$) to absolute privacy protection (when $\epsilon = 0$). DP even allows for different queries in a particular release to receive different amounts of protection. The 2020 Census used this to provide more accuracy for the count of people by race at the geographic level of the United States than at the block level (U.S. Census Bureau 2021)

To get a visceral understanding of tunability's impact on accuracy and privacy loss, recall the example of mixed race households in the introduction. More accuracy means that demographers will likely publish better papers with improved p-values using the published data. It also means that the hoodlums' will be more likely to find mixed race households when they pound on the door.

3. The Costs of Differential Privacy

Just because DP can work with any value of $0 \leq \epsilon \leq \infty$ does not mean that all of these values provide adequate privacy protection. This is why the claim that a data release is performed using DP, without discussing the specific value for ϵ (and ideally the entire DP mechanism), says absolutely nothing about the privacy protection or the accuracy of the resulting data. Organizations using DP need to make hard decisions about the trade-off between the benefit of a specific data release and the privacy loss that the data subjects might experience as a result. Few organizations have the systems or the technical skills to make these decisions, resulting in programming, training and hiring costs.

A second cost comes from the framework's underlying definition of privacy loss as the increase in accuracy of an arbitrary query based on the released data. A data intruder with significant background data will be able to make better use of that data release than a data intruder with no background data. DP thus offers "relative", not absolute, privacy protection. The cost here is social: DP does not allow its practitioners to make statements that they have protected specific confidential data against all possible attackers. This is a cost to the credibility of statistical practitioners who must now modify their previous promise they could broadly protect privacy while publishing highly accurate data; technological progress has broken that promise.

There is a third cost: an organization has made a commitment to using differential privacy may discover that it is unable to do so because the science of differential privacy is not yet sufficiently advanced (Daily 2022).

4. Correcting the Record

Mulalidhar and Domingo-Ferrer's assertion that the 2020 Census used a privacy budget of value of $\epsilon = 19.61$ is correct, but misleading, as Mulalidhar and Domingo-Ferrer fail to note that there is no single query over the 2020 Census confidential data for which this privacy loss value is used. Instead, the privacy loss is spread over tens of millions of individual queries to produce the so-called "noisy measurements". So while it is true that "$\epsilon = 19.61$ translates to binary RR with $p = 0.99999999696$," this statement is not relevant to the use of differential privacy in the 2020 Census.

There has also been some confusion regarding the source of apparent errors in the redistricting dataset. An article in The New York Times stated that DP was responsible for allocating 13 adults and one child to Census Block 1002 in downtown Chicago, a block that "consists entirely of a 700-foot bend in the Chicago River" (Wines 2022). In fact, the TopDown algorithm implements a constraint such that "the number of householders (person one on the questionnaire) cannot be greater than the number of housing units" (Abowd et al. 2022). Hence, the Maser Address File must have had at least one housing unit on that block, as no noise was ever added to the housing unit counts at any level of geography. And as the Census Bureau's Master Address File (MAF) does not allow housing units to be placed underwater, the curiosity of block 1002 was not a result of DP. It was likely either the assignment to that block of floating living quarters by the address canvasser, or an error in the MAF.

Other apparent block-level inconsistencies in the redistricting dataset may have been the result of the TopDown algorithm (Abowd et al. 2022) producing unlinked outputs for the demographic and housing files. For example, there may be a block with a single housing unit and an unrealistically high number of occupants. The purpose of the algorithm was not to produce accurate data for single blocks: it was to produce data for blocks that, when added together, would produce accurate data for larger geographic areas. In this, the Census Bureau was largely successful, as evidenced by numerous data analyses published since that make use of the data.

While it is correct in asserting that reconstruction and reidentification are different, Mulalidhar and Domingo-Ferrer's thought experiment presented in Subsection 5.2, "Misinterpreting reconstruction as reidentification may overstate or understate the reidentification risk," does not accurately present one of the primary risks that DP was designed to protect: the release of confidential information about race.

The commercial data purchased in 2010 did not include self-reported race; in general, such data are not for sale. It is also very difficult to purchase detailed information about children.

Mulalidhar and Domingo-Ferrer seem to argue that the Census Bureau was wrong to adopt differential privacy because of the impact on statistical accuracy, and that it should have used a combination of data suppression and swapping. Mulalidhar and Domingo-Ferrer fail to address a very real harm to the statistics of vulnerable populations done by suppression and swapping. That harm is "erasure". By design, these techniques hide the statistical contribution of minorities. Another problem with swapping is that data users cannot evaluate the impact that swapping has on their statistics in any straightforward way. Both are unacceptable in a representative democracy.

The Census Bureau has started to document the range of reasons why it moved to differential privacy for the 2020 Census. McKenna (2018, 2019) detail how the Census Bureau struggled to protect both the tables and microdata for decennial censuses between 1970 and 2010, periodically making the data releases less accurate in the interest of protecting privacy. This struggle comes from the diversity within the US population: "About 57 percent of the 2010 Census population were 'unique' at the smallest census geography, block level, meaning they were the only people in their block with a specific combination of sex, age (in years), race (any of the 63 possible Office of Management and Budget race combinations), and Hispanic/Latino ethnicity" (McKenna 2018).

On May 25, 2021, the Census Bureau released to the Census Scientific Advisory Committee the results of an experiment of applying the suppression rules from the 1980 Census to two of the proposed data releases for the 2020 Census (using the data from the 2010 Census). Using only primary suppression, it found that 83.8% of the block-level cells in the P3 table (Race for the population 18 years and over), 95.7% of the blockgroup level cells, 84.3% of the tract-level cells, and 51.2% of the county-level cells would have needed to be suppressed. For the P4 table (Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 Years and Over), the suppression numbers are 87.7%, 100.0%, 99.7%, and 84.2% (Hawes 2021a). And primary suppression alone would not have offered sufficient protection: the suppression advocated by Mulalidhar and Domingo-Ferrer would have also required secondary suppression, leading to further erasure, and making the data even more unsuitable for redistricting.

Suppression was not an option for the 2020 Census. What about swapping?

Although the Census Bureau has not released details of the swapping algorithm used for previous decennial censuses in sufficient detail to replicate the algorithm, a review of the scientific literature finds that swapping has unpredictable and uncontrollable impacts on the accuracy of the statistics not explicitly held constant by the swapping algorithm. Swapping creates significant and measurable bias that cannot be controlled.

Fienberg and McIntyre (2005) write: "Work by Aoki and Takemura (2003) and unpublished results of de Loera and Ohn effectively demonstrate the possibility that the existence of nonsimple basis elements can yield multimodal exact distributions or bounds for cells where there are gaps in realizable values." And: "These results suggest that data swapping as originally proposed by Dalenius and Reiss does not generalize in ways that they thought."

Kim (2015) writes: "Data swapping in its simplest form, wherein a fraction of households is swapped at random, will 'normalize' the strengths of the joint distributions of categorical variables, instead of lowering them." And: "This effect is still observed even when a primitive matching stage is included, so that two households may only be swapped if they match on some predefined set of key variables. The further addition of a minimal targeting stage in the data swapping procedure is shown to impact the statistical quality of the data in an inconsistent way: by deciding to implement a generic selection criterion for at-risk households, even the expected direction of swapping's effect on the joint distributions can no longer be predicted."

Contrary to Mulalidhar and Domingo-Ferrer, the Census Bureau has published specific results comparing DP and swapping. In its efforts to analyze the amount of swapping that would have been needed for the 2020 Census in light of the database reconstruction experiment, the Census Bureau performed two experiments: a swap of the 2010 unswapped data at 5% and at 50%. The 50% swap resulted in 70% of the census tracts being perturbed, but only dropped the confirmed reidentification rate from 16.85% to 12.96%. Additional results can be found in Hawes (2021).

5. How Much Noise is Needed?

A recurring idea in Mulalidhar and Domingo-Ferrer is that there is an objectively correct amount of noise required to provide an absolute level of privacy protection. For example, the article states that the principle finding of Dinur and Nissim (2003), that the amount of noise needed to protect a database of size $O(n)$ from a polynomial adversary is $O(\sqrt{n})$, means that roughly 1,000 bits in a 1,000,000 bit database need to differ from their true query to protect privacy. The idea continues when Mulalidhar and Domingo-Ferrer relate the finding of Bach (2021) regarding ϵ-DP.

However, Dinur and Nissim (2003) makes a variety of arguments using not just big-O notation (e.g. $O(\sqrt{n})$ but also little-o notation ($o(\sqrt{n})$) and omega notation ($\Omega(n)$), all in an effort to establish various bounds on the privacy that is achieved by output perturbation. Section 3, "Impossibility Results," states "any database algorithm that is within $o(\sqrt{n})$ perturbation, is non private with respect to polynomial time adversaries. More accurately, we show that whenever the perturbation is smaller than $\sqrt{n}$, a polynomial number of queries can be used to efficiently reconstruct a 'good' approximation of the entire database." This does not mean that $\sqrt{n}$ perturbations are sufficient for dataset privacy: it means that *fewer* than $\sqrt{n}$ perturbations are *not* sufficient—and then, only for the kind of database and queries considered in the article. The weakness of this result is one of the factors that led to the invention of differential privacy.

Meanwhile, Bach's statement is not a comment about whether or not DP overprotects: it is a comment that significant error can be introduced into query results that are protected with the Laplace distribution because of its long tails. This has been known since the invention of DP, and it is one of the reasons that mechanisms implementing formal relaxations of DP. Mechanisms implementing (ϵ, δ)-DP use Gaussian noise and those implementing zero-concentrated DP (z-CDP) use discrete Gaussian noise. The disadvantage is that such mechanisms may occasionally fail to provide the full protective power of ϵ-DP, but they have the advantage of producing more accurate results for any given value of privacy loss (ϵ or ρ) and of allowing characterization of the effects of the entire distribution of the associated noise, not just extrema (Kifer et al. 2022).

Bach's comments about what is and is not adequate to protect confidential data under the law do not refer to the law of the United States, evidenced by its footnote 20: "In fact, the related discussion in the U.S. on 2020 census protection has a legal dimension addressing what exactly is protected by U.S. privacy law (Ruggles et al. 2019; Mervis 2019), which may have been fueled also by a political controversy on the questions to be included in the 2020 U.S. census (Mervis 2018). In any case, all statements here on EU privacy law purely reflect the authors' personal perceptions."

5.1. What Does it Really Mean to "over-protect" or "over-state risk?"

One of the consistent critiques of DP is that it over-protects data and that those who are using it over-state the risks that DP protects against. These complaints ignore the fact that DP is a tunable protection mechanism. DP doesn't dictate how much noise to add: that is a policy question for those who use it.

Likewise, Mulalidhar and Domingo-Ferrer fail to discuss the efficiency of the noise that is added. Like other SDL methods, DP broadly has two modes of operation: the local model and the trusted curator model. In the local model, noise is added to the individual microdata. In the trusted curator model, noise is added to the query results that are output. RR is an example of DP run in local model.

It is less efficient to add noise to input microdata (local model) than to add noise to query results (trusted curator model). With counting queries, the amount of noise added to each record in local model is the same as the total amount of noise added to the final result in the trusted curator model. A query that asks "how many children are on this block" for a block with 1,000 people will require a single draw of noise when run in trusted curator model, but 1,000 draws of noise (with the same parameters) that are then added together when the query is run in local model. The result is more noise, less signal. It is for this reason that Google decided to stop using RR to protect privacy-sensitive statistics reported by the Chrome web browser several years after it developed and deployed the technology (Erlingsson et al. 2014).

While Dwork did recommend in 2011 (Dwork 2011) that appropriate values for ϵ might be 0.01, 0.1, $ln(2)$ or $ln(3)$, today we know a lot more about differential privacy and how to set ϵ than Dwork did in 2011. For example, now we know that RR is mathematically equivalent to differential privacy's local model, and we understand why the trusted curator model is inherently more efficient, although it does require that respondents trust the curator with their confidential data.

Mulalidhar and Domingo-Ferrer state "Taking a very small ϵ entails unaffordable utility loss, but taking ϵ very large entails very little noise addition and offers little to no protection against reconstruction, let alone reidentification." This wrongly implies that there are objectively correct values for data utility and privacy loss.

In fact, choosing a small ϵ means that the data owner has chosen to highly value privacy at a cost of data utility, while large ϵ values means the data owner has chosen to highly value data utility at the cost of privacy.

6. Finding a Constructive Way Forward

It's true that DP is more complex to understand than traditional SDL techniques; the same was true when Diffie and Hellman (1976) first published their discovery of public key cryptography. Today we are 17 years since the invention of DP. But back in 1994, 18 years after the invention of public key cryptography, there was still no widespread consensus that the internet required such a complex and difficult-to-understand privacy protecting technology. That December saw the introduction of Netscape Navigator with its new and incompatible "https:" protocol that made it possible to send credit card numbers over the web securely. Netscape Navigator required upgrading every web server on the planet, but it led to the Internet's commercialization. Even after Navigator was released, it was not until 2015 that Tim Bernes-Lee, the inventor of the World Wide Web, called on us to encrypt every web page (Berners-Lee 2015).

We are likewise in the early days of DP. With DP, we finally have a mathematical approach for modeling the privacy loss that individuals experience with incremental data releases: we can now mathematically model the mosaic effect! The findings of DP have resulted in considerable push-back from some statisticians and data users, but a new generation of scientists are learning to work with them, in line with Kuhn's theory of scientific progress (Kuhn 1962).

SDL experts can help by formalizing their protection goals and assumptions. It would also be useful for the SDL community to release open source software that implements their mechanisms and publish experiments that systematically test protection methods in use today on realistic data.

Our understanding of DP has expanded since 2011. This is the hallmark of scientific progress. What's needed now is to find other opportunities for improving the mathematical rigor of privacy protection throughout our profession.

7. References

Abowd, J.R.A., R. Cumings-Menon, S. Garfinkel, M. Heineck, C. Heiss, R. Johns et al. 2022. "The 2020 Census Disclosure Avoidance System TopDown Algorithm." *Harvard*

Data Science Review 2. Available at: https://hdsr.mitpress.mit.edu/pub/7evz361i (accessed April 2023).

Abowd, J.M., G.L. Benedetto, S.L. Garfinkel, S.A. Dahl, A.N. Dajani, M. Graham, M.B. Hawes et al. 2020. *The modernization of statistical disclosure limitation at the U.S. Census Bureau*, Available at: https://www.census.gov/library/working-papers/2020/adrm/CED-WP-2020-009.html (accessed April 2023).

Bach, F. 2021. "Differential Privacy and Noisy Confidentiality Concepts for European Population Statistics." *Journal of Survey Statistics and Methodology* 10(3): 642–687. DOI: https://doi.org/10.1093/jssam/smab044. eprint: https://academic.oup.com/jssam/article-pdf/10/3/642/44275540/smab044.pdf.

Berners-Lee, T. 2015. *Web Security—TLS Everywhere, not https: URIs* Available at: https://www.w3.org/DesignIssues/Security-NotTheS.html (accessed April 2023).

Chappell, B. 2020. *Texas Supreme Court OKs state child abuse inquiries into the families of trans kids*. Available at: https://www.npr.org/2022/05/13/1098779201/texas-supremecourt-transgender-gender-affirming-child-abuse.

Daily, D. 2022. "Disclosure Avoidance Protections for the American Community Survey." *Random Samplings*. Available at: https://www.census.gov/newsroom/blogs/-random-samplings/2022/12/disclosure-avoidance-protections-acs.html (accessed April 2023).

Denning, D.E., and J. Schlöorer. 1980. "A Fast Procedure for Finding a Tracker in a Statistical Database." *ACM Trans. Database Systems* 5(1): 88–102. DOI: https://doi.org/10.1145/320128.320138.

Dick, T., C. Dwork, M. Kearns, T. Liu, A. Roth, G. Vietri, and Z.S. Wu. 2023. "Confidence-ranked reconstruction of census microdata from published statistics." *Proceedings of the National Academy of Sciences* 120(8): e2218605120. DOI: https://doi.org/10.1073/pnas.2218605120.

Diffie, W., and M. Hellman. 1976. "New directions in cryptography." *IEEE Transactions on Information Theory* 22(6): 644–654. DOI: https://doi.org/10.1109/TIT.1976.1055638.

Dinur, I., and K. Nissim. 2003. "Revealing Information While Preserving Privacy." In Proceedings of the Twenty-second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems: 202–210. PODS '03. June, San Diego, California: ACM. DOI: https://doi.org/10.1145/773153.773173.

Dwork, C. 2011. "A firm foundation for private data analysis." *Communications of the ACM*, 54(1): 86–95. DOI: https://doi.org/10.1145/1866739.1866758.

Dwork, C., F. McSherry, K. Nissim, and A. Smith. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis." In Proceedings of the Third Conference on Theory of Cryptography: 265–284. TCC'06. New York, NY: Springer-Verlag. DOI: https://doi.org/10.1007/11681878_14.

Dwork, C., and A. Roth. 2014. "The Algorithmic Foundations of Differential Privacy." In Foundations and Trends in Theoretical Computer Science, 9: 211–407. 3–4. NOW. DOI: https://doi.org/10.1561/0400000042.

Erlingsson, U., V. Pihur, and A. Korolova. 2014. "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response." In Proceedings of the 2014 ACM SIGSAC

Conference on Computer and Communications Security: 1054–1067. CCS ’14. Scottsdale, Arizona, USA: ACM. DOI: https://doi.org/10.1145/2660267.2660348.

Fienberg, S.E., and J. McIntyre. 2005. “Data Swapping: Variations on a Theme by Dalenius and Reiss.” *Journal of Official Statistics* 21 (2): 309–323. Available at; https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/data-swapping-variations-on-a-theme-by-dalenius-and-reiss.pdf.

Hansen, M. 2018. “To Reduce Privacy Risks, the Census Plans to Report Less Accurate Data.” *The New York Times* Available at: https://www.nytimes.com/2018/12/05/upshot/to-reduce-privacy-risks-the-census-plans-to-report-less-accurate-data.html (accessed April 2023).

Hawes, M. 2021a. *The Census Bureau’s Simulated Reconstruction-Abetted Re-identification Attack on the 2010 Census*. Available: https://www.census .gov/data/academy/webinars/2021/disclosure-avoidance-series/simulated-reconstruction-abetted-re-identification-attackon-the-2010-census.html (accessed April 2023).

Hawes, M. 2021b. *The Census Bureau’s Simulated Reconstruction-Abetted Re-identification Attack on the 2010 Census*. Available at: https://www.census .gov/data/academy/webinars/2021/disclosure-avoidance-series/simulated-reconstruction-abetted-re-identification-attackon-the-2010-census.html (accessed April 2023).

Jarmin, R. 2019. *Census Bureau Adopts Cutting Edge Privacy Protections for 2020 Census*. Available at: https://www.census.gov/newsroom/blogs/random-samplings/2019/02/census bureau adopts.html (accessed April 2023).

Kenthapadi, K., N. Mishra, and K. Nissim. 2013. “Denials leak information: Simulatable auditing.” *Journal of Computer and System Sciences* 79 (8): 1322–1340. DOI: https://doi.org/10.1016/j.jcss.2013.06.004.

Keyes, O., and A.D. Flaxman. 2022. “How Census Data Put Trans Children at Risk.” *Scientific American* Avialable at: https://www.scientificamerican.com/article/how-census-data-put-trans-children-at-risk/ (accessed April 2023).

Kifer, D., J.M. Abowd, R. Ashmead, R. Cumings-Menon, P. Leclerc, A. Machanavajjhala, W. Sexton, and P. Zhuravlev. 2022. *Bayesian and Frequentist Semantics for Common Variations of Differential Privacy: Applications to the 2020 Census*. Technical report, CED Working Paper CED-WP-2022-004. Suitland, MD: Center for Enterprise Dissemination, US Census Burea. Available at: https://www.census.gov/library/working-papers/2022/adrm/CED-WP-2022-004.html (accessed April 2023).

Kim, N. 2015. “The Effect of Data Swapping on Analyses of American Community Survey Data.” *Journal of Privacy and Confidentiality* 7(1). DOI: https://doi.org/10.29012/jpc.v7i1.644.

Kleinberg, J., C. Papadimitriou, and P. Raghavan. 2000. “Auditing Boolean Attributes.” In Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems: 86–91. PODS ’00. Dallas, Texas, USA: ACM. DOI: https://doi.org/10.1145/335168.335210.

Kuhn, T.S. 1962. *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Leclerc, P. 2019. “Results from a Consolidated Database Reconstruction and Intruder Re-Identification Attack on the 2010 Decennial Census.” In Challenges and New Approaches for Protecting Privacy in Federal Statistical Programs: A Workshop. June.

Available at: https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse 193509.pdf (accessed April 2023).

McKenna, L. 2018. *Disclosure Avoidance Techniques Used for the 1970 through 2010 Decennial Censuses of Population and Housing*. Technical report CDAR2018-01. U.S. Census Bureau. Available at: https://www.census.gov/library/working-papers/2018/adrm/ces-wp-18-47.html (accessed April 2023).

McKenna, L. 2019. *Disclosure avoidance techniques used for the 1960 through 2010 Decennial Censuses of Population and Housing Public Use Microdata Samples*. Technical report. U.S. Census Bureau. Available at: https://www2.census.gov/adrm/CED/Papers/CY19/2019-04-McKenna-Six%20Decennial%20Censuses.pdf (accessed April 2023).

Mervis, J. 2018. "Trump Officials Claim They Can Avoid 2020 Census Problems Caused by Controversial Citizenship Question. Experts Are Very Skeptical." *Science* (4). DOI: https://doi.org/10.1126/ science.aat8801.

Mervis, J. 2019. "Can a Set of Equations Keep US Census Data Private?" *Science* (10). DOI: https://doi/10.1126/science.aaw5470.

Muralidhar, K., and J. Domingo-Ferrer. 2023. See article in this issue.

Ruggles, S., C. Fitch, D. Magnuson, and J. Schroeder. 2019. "Differential Privacy and Census Data: Implications for Social and Economic Research," *AEA Papers and Proceedings* 109: 403–408. DOI: https://doi.org/10.1257/pandp.20191107.

Shlomo, N. 2018. "Statistical Disclosure Limitation: New Directions and Challenges." *Journal of Privacy and Confidentiality* 8(1). DOI: https://doi.org/10.29012/jpc.684. https://journalprivacyconfidentiality.org/index.php/jpc/article/view/684.

U.S. Census Bureau. 2021. *Disclosure Avoidance for the 2020 Census: An Introduction*. Available at: https://www.census.gov/library/publications/2021/decennial/2020-census-disclosureavoidance-handbook.html.

Wines, M. 2022. "The 2020 Census Suggests That People Live Underwater. There's a Reason." *The New York Times* Available at: https://www.nytimes.com/2022/04/21/us/-census-data-privacy-concerns.html (accessed April 2023).

Received January 2023
Revised April 2023
Accepted May 2023

Journal of Official Statistics, Vol. 39, No. 3, 2023, pp. 411–420, http://dx.doi.org/10.2478/JOS-2023-0019

A Rejoinder to Garfinkel (2023) – Legacy Statistical Disclosure Limitation Techniques for Protecting 2020 Decennial US Census: Still a Viable Option

Krishnamurty Muralidhar[1] *and Josep Domingo-Ferrer*[2]

In our article "Database Reconstruction Is Not So Easy and Is Different from Reidentification", we show that reconstruction can be averted by properly using traditional statistical disclosure control (SDC) techniques, also sometimes called legacy statistical disclosure limitation (SDL) techniques. Furthermore, we also point out that, even if reconstruction can be performed, it does not imply reidentification. Hence, the risk of reconstruction does not seem to warrant replacing traditional SDC techniques with differential privacy (DP) based protection. In "Legacy Statistical Disclosure Limitation Techniques Were Not an Option for the 2020 US Census of Population and Housing", by Simson Garfinkel, the author insists that the 2020 Census move to DP was justified. In our view, this latter article contains some misconceptions that we identify and discuss in some detail below. Consequently, we stand by the arguments given in "Database Reconstruction Is Not So Easy. . .".

1. Introduction

In this article, we address some of the issues that Garfinkel (2023) raises regarding our earlier article Muralidhar and Domingo-Ferrer (2023). Before we do however, we feel compelled to make the following preliminary remarks:

(1) Garfinkel makes the following statements:

> "Even if more than half of the matches are incorrect, and an external attacker has no means of confirming them" (Jarmin 2019), a group of hoodlums out to terrorize mixed race families can easily confirm the putative matches by visiting the addresses with the intent to commit acts of violence.
>
> More accuracy means that demographers will likely publish better papers with improved p values using the published data. It also means that the hoodlums will be more likely to find mixed race households when they pound on the door.

There is no evidence to support the above claims. Using real data from the 2010 Census, Muralidhar (2022, Table 11, 321) has shown that the reconstruction procedure assigns race and ethnicity randomly. Specifically, due to the uncertainty in the reconstruction of age, he showed that, for a given (sex, age-group) every individual in the block could potentially be assigned every (race, ethnicity) combination present for that (sex, age-group). Hence, the putative identification is essentially random and is useless in assigning the correct (race,

[1] University of Oklahoma, Price College of Business, Dept. of Marketing and Supply Chain Management, 307 West Brooks, Adams Hall Room 10, Norman, OK 73019, U.S.A. Email: krishm@ou.edu

[2] Universitat Rovira i Virgili, Dept. of Computer Engineering and Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia, Av. Països Catalans 26, 43007 Tarragona, Catalonia, Spain. Email: josep.domingo@urv.cat

ethnicity) to a particular individual. The adversary will have to visit every household. Hence, this claim is baseless fearmongering intended to inflame rather than inform.

(2) The author makes another similar statement relating to trans children in Texas, extracted from a non-refereed Opinion in the Scientific American (Keyes and Flaxman 2022), where the authors claim that they can identify children who changed gender between 2010 and 2020. It is important to note that the results are based on simulated data and not on real data. Perhaps more importantly, this claim of reidentification is not based on identifying individuals; it is based exclusively on comparing unique records within a given geography whose attributes matched on (age-adjusted-by-ten-years, race, ethnicity), but differed on gender (Flaxman and Keyes 2022). But there are many reasons that may explain such a difference, the most common being that households had moved between the two censuses. In fact, even for their own simulated data, the authors find that, of the 70,184 individuals identified, only 657 (0.94%) were trans children, while the remaining 69,527 (99.06%) individuals were incorrectly identified as trans children (Flaxman and Keyes 2022, 11). Note that this verification was only possible only because the data were simulated, and the authors knew the correct identity of the trans children. Without this information, there is no way to establish the identity of the trans children. Hence, there is no basis to the claim that trans children can be identified. Again, baseless fearmongering intended to inflame rather than inform.

We now proceed to address the main content of Garfinkel, much of which is only tangentially relevant to Muralidhar and Domingo-Ferrer. Sections 1, 2, and 3 in Garfinkel are a primer on differential privacy; Section 6 offers a defense of differential privacy. Garfinkel engages in an extensive discussion of suppression which we never addressed in our article. We ignore this discussion as well. We focus our attention primarily on Sections 4 and 5 in Garfinkel and, where appropriate, on specific comments in the other sections.

To facilitate the readability of this article, we first summarize the specific issues raised by Muralidhar and Domingo-Ferrer:

(1) The relevance of Dinur and Nissim (2003) in Census database reconstruction.
(2) The Census reconstruction procedure.
(3) The Census reidentification procedure.
(4) The impact of geography and policy choices on the performance of swapping and the DP-based method.
(5) A discussion of the privacy offered by the DP-based method; and
(6) A discussion of the accuracy of the data generated using the DP-based method.

In the following sections, we address each issue above in greater detail in view of Garfinkel.

2. The Dinur-Nissim Reconstruction

One of the primary motivations of our study was to highlight the fact that, while the Dinur and Nissim (2003) study is an important contribution, it has limited relevance in the context of the Census. Specifically, we wanted to bring attention to the "CD Model" described in Dinur and Nissim (2003), which states that a data curator may be able to modify the original data set to produce a "modified" data set that could possibly be securely shared. It is easy to see that this is exactly the procedure that the U.S. Census Bureau (Census Bureau) employs. Statistical disclosure limitation is applied to the Census Edited File to create the Hundred-percent Detail File. All tabular data are then generated from the Hundred-percent Detail File (and no access is provided to the Census Edited File) (Zayatz et al. 2009). Furthermore, unlike the Dinur and Nissim (2003) reconstruction procedure where the adversary can issue any query, in the Census context only the results of a pre-determined set of queries (the tabular data results) are allowed. The result of this is that the adversary can only reconstruct the Hundred-percent Detail File and never the Census Edited File. Another possible way to state this is as follows: you can only reconstruct what the Census allows you to reconstruct and nothing else.

Garfinkel contends that there are two types of protection models, local and trusted curator. According to him, the local model involves adding noise to microdata while the trusted curator model involves adding noise to the output. But this is a misunderstanding of the methods. Indeed, the local and trusted curator models represent *where* the protection is applied, not what protection is applied. In a local model, the respondent modifies the response value prior to sharing with the curator. In a trusted curator model, the respondent shares the true value with the curator. In fact, it is possible to implement noise addition to microdata in the trusted curator model (which is the case with the American Community Survey (McKenna 2019). Note that, by definition, a local model precludes adding noise to the output.

Garfinkel also questions our statements regarding the level of noise required to prevent reconstruction. Our statements are based directly on the Dinur and Nissim (2003) conclusions (our emphasis): "a database algorithm that is within $O(\sqrt{n})$ perturbation and is private against polynomial adversaries in the strongest possible sense." Thus, our original statement is accurate. That there are other bounds is irrelevant.

Garfinkel also makes another dubious comment that adding noise at the microdata level is less efficient than adding noise at the query level: "With counting queries, the amount of noise added to each record in local model is the same as the total amount of noise added to the final result in the trusted curator model." This claim is made without any substantiation or support. We already addressed this exact question in our article (Muralidhar and Domingo-Ferrer 2023, subsec. 3.5) and showed that when many queries are issued, the noise added by a DP-based method will be higher than with an alternate procedure such as randomized response. The statement of Bach (2022) is consistent with our analysis.

Finally, the 2010 Decennial Census tabular data release did not involve any noise addition. Instead, data swapping was used as the disclosure limitation tool (Zayatz et al. 2009). One of the biggest benefits of data swapping is that it does not modify the original values of each attribute. In addition, data swapping was implemented with the explicit requirement that voting age population must be preserved exactly at the block level. As a result, the question how many children are in this block will be answered exactly, without any error.

3. The Census Reconstruction Procedure

The entire justification for adopting differential privacy for the 2020 Census was based on the results of the reconstruction and reidentification experiments of the 2010 tabular data release. The Census Bureau has also made the argument explicit in Abowd (2021):

> "... only over the last few years have computing power and the sophisticated numerical optimization software necessary to perform these types of reconstructions advanced enough to permit reconstruction attacks at any significant scale."

Ruggles (2018) has argued from the very beginning that the Census Bureau reconstruction amounts to little more than creating a list of individuals from the tabular data: "any tabular data can be expressed as microdata." This was proved by Muralidhar (2022), who was able to perform the reconstruction using only the data released to the public and an Excel spreadsheet on a laptop, without any numerical optimization software. He observed that, at the *tract* level, the Census Bureau provides the count of (Sex × Age × Race × Ethnicity). Similar data are provided at the *block* level, but with predefined Age Bins in place of individual year of age. Muralidhar (2022) showed that the data at the block level can be reconstructed by randomly assigning the counts of the individual year of age at the tract level to the corresponding Age Bins at the block level. This essentially proves the claim by Ruggles (2018) that the Census tabular data can be expressed as microdata, albeit with a little arithmetic. More importantly, it disproves the claim by Abowd (2021) that extensive computing power and numerical optimization is necessary to perform the reconstruction.

In 2022, the Census Bureau even stopped reconstructing individual years of age. Instead, reconstruction was performed using only Age Bins. With this change, reconstruction is little more than creating a list based on the cell counts from the tabular data, which goes to further support Ruggles (2018) and contradict Abowd (2021).

Note that Garfinkel does not acknowledge Muralidhar (2022) and the detailed explanation the latter provides as to how the reconstruction is simple and straightforward and that the technology needed to perform reconstruction has been available for 50+ years.

An even more important contribution of Muralidhar (2022) is that he showed that, even for small geography (at the Census tract level), there are billions of feasible reconstructions that are equally likely. This has significant implications for the claims of reidentification as we now discuss.

4. The Impact of Policy and Geography Choices

Of all the claims made by the Census Bureau, the most dramatic is the following (Abowd 2021, app B):

> "If the external data on name, address, sex, and age are comparable to the 2010 Census, then the attacker will putatively re-identify 238 million persons (77% of the 2010 Census resident U.S. population). Confirmed re-identifications will be 179 million (58% of the same population). This means that with the best quality external data, relative to the 2010 Census, as many as 179 million persons could be correctly re-identified using the attack strategy outlined here."

Abowd (2021) leaves no doubt as to what he means by reidentification:

> ". . . link the reconstructed microdata to a real name and address associated with the block, sex, age, race, and ethnicity on the reconstructed microdata."

These are dramatic statements because they directly contradict the legal requirements in Title 13 of the United States Code, described in Abowd et al. (2020) as follows:

> "Additionally, the Department of Commerce (2017), in which the Census Bureau is housed, has issued directives regarding the protection of personally identifiable information (PII) and business identifiable information (BII). These directives largely mirror those issued by other government agencies and prohibit release of information that can be used "to distinguish or trace an individual's identity, such as their name, social security number, biometric records, etc., alone or when combined with other personal or identifying information which is linked or linkable to a specific individual, such as date and place of birth, mother's maiden name, etc."

Abowd (2021) states explicitly that the primary reason for adopting differential privacy for the 2020 Decennial Census was that the tabular data release from the 2010 Decennial Census violated Title 13 requirements:

> "To defend against this known vulnerability, the Census Bureau explored different confidentiality methods that explicitly defend against database reconstruction attacks and concluded that the best tool to protect against this modern attack while also preserving the accuracy and usability of data products comes from the body of scientific work called "differential privacy"."

Muralidhar and Domingo-Ferrer pointed out that the claims of reidentification were vastly overstated. These results are supported by the empirical analysis of 2010 Census tabular data by Muralidhar (2022). The reconstruction procedure in Abowd (2021) was shown to directly contradict the reidentification procedure suggested by the Census Bureau's own Research and Methodology Directorate (McKenna 2019). The claims of reidentification also contradict Dr. Ron Jarmin, Acting Director of the Census Bureau, who said: "more than half the matches are incorrect." (Jarmin 2019) Based on all this information, we were pointing out the fact that what Abowd (2021) claims as reidentification was actually the accuracy of reconstruction, as illustrated in Subsection 5.2. of Muralidhar and Domingo-Ferrer. What is lacking is the ability to uniquely link the respondents in the reconstructed data to their identity (see Subsection 5.2 in Muralidhar and Domingo-Ferrer for more details).

Garfinkel claims that "there are many ways that even unidentified census microdata with block-level geography can cause harm." This may be true, but it is a far cry from claiming that the reconstructed data can be linked to individual names and addresses. Garfinkel now claims that even if the reidentification was incorrect, it can be verified by the adversary by visiting the address. If we take this to its logical conclusion, for the adversary to make the same claim as Abowd (2021) regarding reidentification, the adversary will have to conduct a new Census.

In summary, Garfinkel offers nothing to contradict our conclusion that the reidentification claims regarding the 2010 Census tabular data release have been vastly overstated by the Census Bureau reconstruction attacks.

5. The Impact of Policy and Geography Choices

Muralidhar and Domingo-Ferrer pointed out that the performance of data swapping and the DP-based method were evaluated using different benchmarks. Perhaps the most important of this is reflected in the *invariants*, that is, counts that are preserved and the geographic level at which these counts are preserved. The 2010 Decennial Census tabular data release preserved exact counts at the block level for the total population, population aged 18 and above, and population below age 18. Regarding the 2020 Decennial Census, a 2017 Census publication (Dajani et al. 2017) states that:

> By agreement with the Department of Justice (2000), the Census Bureau will provide exact counts at the Census block level for the following variables:
>
> - Number of people: total, age 18+ (voting age), and less than age 18,

The updated version of the same document in 2020 (Abowd et al. 2020) states:

> The Census Bureau currently plans the following invariants for the 2020 Census data publications:
>
> - Total number of people by state;

This is a remarkable change in policy (and one that has had a significant detrimental impact on data accuracy as we will discuss later). While the 2017 invariants would have been consistent with the 2010 invariants, the 2020 invariants have remarkably relaxed the requirement, from the block level to the state level. From the Census demonstration data products released to the public, we know that the DP-based method for the demonstration products only preserved totals at the state level, while the original 2010 tabular data release using swapping preserved totals at the block level. Considering that the number of blocks varies from 6,012 (Washington D.C.) to 668,757 (Texas), this represents a massive change in policy. Muralidhar and Domingo-Ferrer point out that such changes in policy have a direct impact on the protection afforded by the statistical disclosure limitation procedure.

It is true that the Census Bureau has recently released comparisons of applying data swapping with different parameters (the proportion of swapped records) (Hawes 2022). However, it is still not clear whether these new parameters were applied with the old policy (invariants at the block level) or the new policy (invariants at the state level). Just changing the proportion of records swapped without changing the invariant requirement is unlikely to change the performance of the swapping algorithm. It would have been relatively easy for the Census Bureau to release data where both data swapping and the DP-based method were held to the same standard. By not doing so, the Census Bureau has made any meaningful comparison impossible. Unfortunately, Garfinkel does not address this issue at all.

Instead, Garfinkel chooses to make new policy by stating “the primary risks that DP was designed to protect: the release of confidential information about race.” This is quite a remarkable statement and, in some ways, a curious one. Garfinkel claims that information regarding race and ethnicity data are “in general, not for sale.” While this claim is arguable, let us assume it is true. If this is the case, then race and ethnicity are the confidential variables that the adversary does not know. Remember that reconstruction was performed by matching age and sex from the reconstructed data to the externally obtained data. This is a common phenomenon encountered in statistical disclosure limitation where the attributes used to perform the matching are referred to as quasi-identifiers. To prevent the adversary from using the quasi-identifiers (age and sex) to learn the confidential attributes (race and ethnicity), privacy models such as k-anonymity

(Samarati 2001) would modify the quasi-identifiers and not the confidential attributes. Thus, if the objective is to prevent reconstruction, it would make more sense to protect the age and sex attributes rather than the race and ethnicity attributes. The rationale is simple – if you cannot reconstruct the data accurately, you cannot infer the confidential attributes.

6. The Privacy Provided by DP

We must admit that we were astounded by the statement in Garfinkel that the privacy loss parameter ϵ and the corresponding randomized response probability is "not relevant to the use of differential privacy for the Census." Garfinkel contends that "there is no single query over the 2020 Census confidential data for which this privacy loss value is used. Instead, the privacy loss is spread over tens of millions of individual queries to produce the so-called noisy measurements."

According to Dwork et al. (2019) Differential privacy allows us to quantify cumulative privacy loss as data are analyzed and re-analyzed, shared, and linked. These differentially private systems will, for the first time, allow us to measure and compare the total privacy loss due to these personal data-intensive activities.

The privacy parameter ϵ represents this total (cumulative) privacy loss. It would be a profound misinterpretation to claim that, since the privacy loss is distributed over tens of millions of individual queries, the cumulative privacy loss does not matter.

Muralidhar and Domingo-Ferrer does not say that "there are objectively correct values for data utility and privacy loss," as Garfinkel asserts. In our opinion, the statements "Taking a very small ϵ entails unaffordable utility loss, but taking very large ϵ entails very little noise addition and offers little to no protection against reconstruction" (Muralidhar and Domingo-Ferrer 2023) and "choosing a small ϵ means that the data owner has chosen to highly value privacy at a cost of data utility, while large ϵ values mean the data owner has chosen to highly value data utility at the cost of privacy" (Garfinkel 2023), say the same thing albeit in different words.

7. Data Accuracy of the DP-based Method

The lack of accuracy of the DP-based method is well documented, as shown in Muralidhar and Domingo-Ferrer. Unfortunately, Garfinkel does not address any of the multiple research studies cited in Muralidhar and Domingo-Ferrer. He only affirms that "the Census Bureau was largely successful, as evidenced by numerous data analyses published since that make use of the data." According to Menger (2021), as many as 11% of the blocks have "impossibilities" (more households than people, zero housing units with non-zero population, zero population but non-zero occupied housing units, children but no adults, and very large households). Whether this constitutes a large successful implementation is up to question. As in many other instances, Garfinkel makes the statement of numerous data analyses, but provides no citation.

Garfinkel has chosen to address one humorous example highlighted in the NY Times of people living under water. He categorically states that this error was not due to DP. But if this were the case, one would have expected the Census Bureau to issue some type of formal explanation for this. Garfinkel further contends that this error could not have happened since, according to Abowd et al. (2022), there is a constraint that "the number of householders (person one on the questionnaire) cannot be greater than the number of housing units. But there have been documented instances where 4.83% [of blocks] have zero occupied housing units which have people living in them (that is, there are no occupied dwellings, but there are people living there)" (Menger 2021). Clearly, the constraint described by Abowd et al. (2022) is being violated.

Finally, we address the rather serious claim by Garfinkel: "MDF fails to address a very real harm to the statistics of vulnerable populations done by suppression and swapping. That harm is erasure. By design, these techniques hide the statistical contribution of minorities." As we have emphasized before, we absolutely deny that we recommend suppression. Removing suppression from the discussion and focusing only on swapping, we can categorically state that data swapping does not result in erasure. Data swapping simply exchanges the attribute values between records. As such, there is no addition or subtraction to the counts. Hence, we reiterate that data swapping does not result in erasure.

The real irony is that the DP-based method does result in erasure. With the DP-based method, the frequency counts are modified by adding or subtracting a random noise term. As a result, every time negative noise is applied to a count of minorities, there is erasure. And there is no way to prevent this from happening. Thus, contrary to the claim made by Garfinkel, it is the DP-based method that is likely to erase the contribution of the minorities, not data swapping.

8. Conclusions

We have carefully gone through the article of Garfinkel. The article offers no direct substantive, justifiable criticism of Muralidhar and Domingo-Ferrer. We have gone to great lengths to address every issue that we found relevant.

Garfinkel calls for finding a constructive way forward. The first step in moving forward is to perform a comprehensive comparison of data swapping and the DP-based method using the same criteria as in the 2010 tabular data release. The second step is to re-evaluate the reidentification claims. Moving forward also requires acknowledging the work of researchers on the other side and not ignoring all their arguments.

9. References

Abowd, J.M. 2021. *Declaration of John M. Abowd. Case no. 3:21-CV-211-RAH-ECM-KCN*, U.S. District Court for the Middle District of Alabama. Available at: https://censusproject.files.wordpress.com/2021/04/2021.04.13-abowd-declaration-alabama-v.-commerce-ii-final-signed.pdf.

Abowd, J., R. Ashmead, R. Cumings-Menon, S. Garfinkel, M. Heineck, C. Heiss, and R. Johns. 2022. "The 2020 Census Disclosure Avoidance System TopDown Algorithm." *Harvard Data Science Review 2*. Available at: https://hdsr.mitpress.mit.edu/pub/7evz361i (accessed May 2023).

Abowd, J.M., G.L. Benedetto, S.L. Garfinkel, S.A. Dahl, A.N. Dajani, M. Graham, and M.B. Hawes. 2020. *The Modernization of Statistical Disclosure Limitation at the U.S. Census Bureau.* Available at: https://www.census.gov/library/working-papers/2020/adrm/CED-WP-2020-009.html (accessed May 2023).

Bach, F. 2022. “Differential Privacy and Noisy Confidentiality Concepts for European Population Statistics.” *Journal of Survey Statistics and Methodology*, 10: 642–687. DOI: https://doi.org/10.1093/jssam/smab044.

Dajani, A.N., A.D. Lauger, P.E. Singer, D. Kifer, J.P. Reiter, A., Machanavajjhala, S.L. Garfinkel, S.A. Dahl, M. Graham, V. Karwa, H. Kim, P. Leclerc, I.M. Schmutte, W.N. Sexton, L. Vilhuber, and J.M. Abowd. 2017. “The modernization of statistical disclosure limitation at the U.S. Census Bureau.” In: Census Scientific Advisory Committee Meeting, September 14–15, Suitland MD, USA. Available at: https://www.census.gov/library/video/2017/2017-09-sac.html (accessed May 2023).

Dinur, I., and Nissim. 2003. “Revealing Information While Preserving Privacy.” In: Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems: 202-210, June 9–12, 2003, San Diego CA, USA. DOI: https://doi.org/10.1145/773153.773173.

Dwork, C., N. Kohli, and D. Mulligan. 2019. “Differential Privacy in Practice: Expose Your Epsilons!” *Journal of Privacy and Confidentiality* 9(2). DOI: https://doi.org/10.29012/jpc.689.

Flaxman, A.D., and O. Keyes. 2022. “The Risk of Linked Census Data to Transgender Children: A Simulation Study,” In: 2020 Census Data Products: Workshop on the Demographic and Housing Characteristics Files, June 21–22, Washington D.C. Available at: https://www.nationalacademies.org/event/06-21-2022/docs/DF3CD148E8CAEC0B93FB88C4A7820AC2D54503ACDD18?noSaveAs = 1 (accessed June 2022).

Garfinkel, S. 2023. “Legacy Statistical Disclosure Limitation Techniques Were Not An Option for the 2020 US Census of Population and Housing.” *Journal of Official Statistics*, this issue.

Hawes, M. 2022. “Reconstruction and Reidentification of the Demographic and Housing Characteristics File (DHC)”, Presentation to the Census Scientific Advisory Committee, September 29–30, 2022, Washington D.C. Available at: https://www2.census.gov/about/partners/cac/sac/meetings/2022-09/presentation-reconstruction-and-re-dentification-of-dhc-file.pdf (accessed March 2023).

Jarmin, R. 2019. *Census Bureau Adopts Cutting Edge Privacy Protection for 2020 Census*. Available at: https://www.census.gov/newsroom/blogs/random-samplings/2019/02/-census_bureau_adopts.html (accessed May 2023).

Keyes, O., and A.D. Flaxman. 2022. “How Census Data Put Trans Children at Risk.” *Scientific American* 21. Available at: https://www.scientificamerican.com/article/how-census-data-put-trans-children-at-risk/ (accessed May 2023).

McKenna, L. 2019. *Disclosure Avoidance Techniques Used for the 1960 Through 2010 Decennial Censuses of Population and Housing Public Use Microdata Samples Research and Methodology Directorate*, U.S. Census Bureau. Available at: https://www.census.gov/library/working-papers/2019/adrm/CED-WP-2019-007.html (accessed May 2023).

Menger, G. 2021. “Using 2020 Census Data.” *Applied Geographic Solutions*. Available at: https://appliedgeographic.com/2021/09/using-2020-census-data/ (accessed May 2023).

Muralidhar, K., J. Domingo-Ferrer. 2023. “Database Reconstruction Is Not So Easy and Is Different from Reidentification.” *Journal of Official Statistics*, this issue.

McKenna, L. 2019. *U.S. Census Bureau Reidentification Techniques*. Research and Methodology Directorate, U.S. Census Bureau. Available at: https://www2.census.gov/adrm/CED/Papers/CY19/2019-04-Reidentification%20studies-20210331FinRed.pdf (accessed May 2023).

Muralidhar, K. 2022. "A Re-Examination of the Census Bureau Reconstruction and Reidentification Attack." In *Privacy in Statistical Databases – PSD 2022*, Paris, France. Springer: 312–323. DOI: https://doi.org/10.1007/978-3-031-13945-1.

Ruggles, S. 2018. *Implications of Differential Privacy for Census Bureau Data Dissemination*, Institute for Social Research and Data Innovation, University of Minnesota. Available at: https://apps.bea.gov/fesac/meetings/Ruggles%20Presentation%20Revised.pdf (accessed May 2023).

Samarati, P. 2001. "Protecting Respondents Identities in Microdata Release." *IEEE Transactions on Knowledge and Data Engineering* 13(6): 1010–1027.

Zayatz, L., J. Lucero, P. Massell, A. Ramanayake. 2009. *Disclosure Avoidance for Census 2010 and American Community Survey Five-year Tabular Data Products*, Census Bureau Research Report RRS2009-10. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2009/adrm/rrs2009-10.pdf (accessed May 2023).

Received May 2023
Revised June 2023
Accepted June 2023

Journal of Official Statistics, Vol. 39, No. 3, 2023, pp. 421–433, http://dx.doi.org/10.2478/JOS-2023-0020

A Note on the Optimum Allocation of Resources to Follow up Unit Nonrespondents in Probability Surveys

Siu-Ming Tam[1], *Anders Holmberg*[1], *and Summer Wang*[1]

Common practice to address nonresponse in probability surveys in National Statistical Offices is to follow up every non respondent with a view to lifting response rates. As response rate is an insufficient indicator of data quality, it is argued that one should follow up non respondents with a view to reducing the mean squared error (MSE) of the estimator of the variable of interest. In this article, we propose a method to allocate the nonresponse follow-up resources in such a way as to minimise the MSE under a quasi-randomisation framework. An example to illustrate the method using the 2018/19 Rural Environment and Agricultural Commodities Survey from the Australian Bureau of Statistics is provided.

Key words: Data quality; nonresponse follow-up; propensity score; weighting adjustment.

1. Introduction

Nonresponse is unavoidable and has become an increasingly challenging issue worldwide for survey practitioners. Item or unit nonresponse, unless properly adjusted for in estimation, will usually have a negative impact on survey data quality (Groves 2006). An extensive literature on nonresponse adjustment exists in journals and books and will not be repeated here (see, for example, Bethlehem 1998; Groves and Couper 1998; Hedlin 2020; Kim and Kim 2014; Kim and Shao 2014; Little 1986; Little and Rubin 2019; Oh and Scheuren 1983; Särndal et al. 1992; Särndal and Lundström 2005; Schouten et al. 2011; Sikov 2018, to name just a few.)

Recent literature on survey design addresses the nonresponse focusses on adaptive designs (see, for example, Groves and Heeringa 2006). Adaptive designs use survey para data and auxiliary data to guide changes to the procedures during data collection to minimise the unit cost of collection whilst addressing nonresponse. Quoting Beaumont et al. (2014), Neusy et al. (2022) argued that adaptive collection procedures, such as call prioritization, cannot reduce the nonresponse bias to a greater extent than a proper nonresponse weight adjustment. We agree with this view. Accordingly, we will not consider adaptive designs further in this article.

In this article, we examine a particular aspect of mitigating the effects of unit nonresponse. We address a classic trade-off for official statistics between cost and quality when collecting data, namely, the optimal deployment of resources earmarked by the survey statistician for nonresponse follow-up (NFU), to meet the NFU budget and maximise data quality. As far as we are aware, this topic has not been addressed in the literature. For example, whilst Neusy et

[1] Australian Bureau of Statistics, Canberra, Australia. 45, Benjamin Way, Belconnen, ACT 2617, Australia. Emails: Stattam@gmail.com, Anders.Holmberg@abs.gov.au and Summer.Wang@abs.gov.au

Acknowledgments: We would like to thank the referees and the Associate Editor for their helpful comments on an earlier version of this article. The views expressed in this article are those of the authors and do not necessarily reflect those of the Australian Bureau of Statistics.

al. (2022) provided a method to determine the NFU sample size, its main aim is to ensure that the NFU budget is not exceeded, and did not, as a twin objective, seek out to maximise the quality of the statistics as a result of the NFU resources deployment.

Currently, a common NFU practice in probability surveys, including those carried out in national statistical offices, is to follow up every nonrespondent with the ultimate aim of maximising the response rate. It is well known that response rates are not necessarily good indicators of survey quality (Curtin et al. 2000; Groves et al. 2004; Groves 2006; Groves and Peytcheva 2006). Indeed, Schouten et al. (2009) gave an example of a Dutch survey in which an increase the response rate of 12.5% points led to an increase of bias of between 0.1 to 0.5% points.

It is therefore clear that increasing response rate in NFU is not the main game. The main game should be to reduce the mean squared error of the estimator of the finite population parameter of interest, subject to the constraint of the NFU budget. This is the main purpose of this article.

The approach we advocate for the allocation of NFU resources to maximise data quality is not the same as taking a random sample of the non-respondents for follow up (see, for example, Elliott et al. 2000; Hansen and Hurwitz 1946; Särndal et al. 1992, 566; Neusy et al. 2022). Our approach allows all nonrespondents to be followed up, but the number of visits to a nonrespondent is determined by where it is located in the RHG, Response Homogeneity Group (Särndal et al. 1992, 578) and the unit cost for NFU in that Group. A challenge with subsampling of the nonrespondents approach is that it requires the assumption of full response in the follow- up phase which is not easy to achieve. On the other hand, the quasi-randomisation framework used in the estimation (Oh and Scheuren 1983) accepts nonresponse as an outcome in the follow-up phase and accounted for it by using weights computed by the inverse of the estimated propensity scores. We use this framework in the article for nonresponse mitigation.

This weighting approach, on the other hand, requires the assumption that (1) every nonrespondent has a non-zero probability to respond and (2) the response mechanism is missing at random. If there is "power of compulsion" enshrined in the statistics legislation, and if the national statistical office uses this power to conduct the survey, there is a good chance that the first assumption is fulfilled. Where the model for the missing-not-at-random (MNAR) nonresponse mechanism is correctly specified, Kim and Morikawa (2022), using an empirical likelihood method and a result from Pfeffermann and Sverchkov (1999)-refer to Note 4 below–showed that the inverse propensity weights (IPWs) can be calibrated to adjust for nonresponse bias and benchmark constraints. In a lecture presented to the 2022 Summer School in Ottawa, Canada, on Modern Methods in Survey Sampling, Professor Jae Kim extended the idea of Kim and Morikawa (2022) to multiple MNAR nonresponse mechanisms, thus reducing the reliance of a particular nonresponse model for the Kim and Morikawa (2022) approach to work. He showed that as long as one of the multiple NMAR models is correctly specified, the estimator is consistent.

In this article, we assume that the nonresponse mechanism is missing at random within the RHGs, Response Homogenous Groups (Särndal et al. 1992, 578). Using this assumption, we propose a method, similar to the approach used for Neyman allocation, to allocate the nonresponse follow-up resources with a view to minimising the mean squared error of the estimator of the population total.

2. Notation and some Well-Known Results

We assume we have a probability sample, s, drawn from a population U of size N with known first order, π_i, and second order, π_{ij}, inclusion probabilities for $i,j \in s$, with $\pi_{ii}{=}\pi_i$. We define the sample weights, d_i, by $d_i = \frac{1}{\pi_i}$. Associated with each unit in U is a target variable of interest, y_i, and a vector of auxiliary variables, $\boldsymbol{x}_i$. We are interested in estimating the population total $T = \sum\limits_{i \in U} y_i$. Due to unit nonresponse, we only have a responding sample, $s_r, s_r \subset s$. Following Oh and Scheuren (1983), we model the nonresponse process using a quasi- randomisation (QR) model, where the responding units are considered to be selected in two stages with the first stage selected from U by probability sampling characterised by known $\pi_i{}'s$ and $\pi_{ij}{}'s$, and the second stage selected from s by a missing-at-random nonresponse process with the probability of responding, also referred to as an (unobserved) propensity score, $\rho_i, i \in s$, defined by $\rho_i = \Pr(\delta_i = 1|\boldsymbol{x}_i$, $i \in s : \boldsymbol{\alpha})$, where δ_i is the response indicator and $\boldsymbol{\alpha}$ is vector of unknown parameters. In practice, ρ_i is estimated by $\hat{\rho}_i = \Pr(\delta_i = 1|\, \boldsymbol{x}_i,\, i \ \in \ s\colon \hat{\boldsymbol{\alpha}})$ where $\hat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$ by assuming a functional form of the probability distribution of ρ, for example, $\frac{\rho}{1-\rho}$ is modelled by logistic regression; or by using non-parametric methods, such as random forest (Breiman 2001).

When using weights to mitigate the effects of nonresponse during estimation, it is customary in national statistical offices to use RHGs for nonresponse adjustment. A RHG is one where the estimated $\rho_i{}'s$ of its group members are the same, that is, missing at random within the RHGs. In reality, they are not the same but similar estimated $\rho_i{}'s$ can be grouped together to form nonresponse adjustment "cells". Little (1986) described this approach as response propensity stratification.

Suppose we partition the sample units by their estimated propensity scores into H RHGs, with the h^{th} group denoted by s_h. Then $s = s_1 \cup .. \cup s_h .. \cup s_H$. Let n_h and m_h denote the size of the sample and the responding units in s_h respectively. Let ρ_{hi} denote the response propensity for the i^{th} unit in the h^{th} RHG. Under the assumption that $\rho_{hi} = \rho_h$, where ρ_h is a constant for every $i \in s_h$, we note that $n'_h = E\big(\sum\limits_{i \in s_h} \frac{\delta_{hi}}{\rho_{hi}} |s\big)$. Using the actual sample size, n, as a plug-in estimate of n'_h, we have $n_h \simeq \sum\limits_{i \in s_h} \frac{\delta_{hi}}{\rho_{hi}} = \frac{m_h}{\rho_h}$ thus giving an estimate of $\rho_h, \hat{\rho}_h = \frac{m_h}{n_h}$, which is just the response rate for the h^{th} RHG. Likewise, let $\rho_{ij} = \Pr(\delta_i = \delta_j = 1s)$. Assuming that $Cov(\delta_{hi}, \delta_{hj})$ is a constant and $Cov(\delta_{hi}, \delta_{h'j}) = 0$ for $i \in h, j \in h'$ and $h \neq h'$, we can likewise show that $\hat{\rho}_{ij} = \frac{m_h(m_h-1)}{n_h(n_h-1)}$ for $i,j \in h, i \neq j$ and $\hat{\rho}_{ij} = \frac{m_h}{n_h}\frac{m_{h'}}{n_{h'}}$ for $i \in h, j \in h'$ and $h \neq h'$. In the sequel, we let $\pi_{hi}, \pi_{hij}, d_{hi}$ and $\boldsymbol{x}_{hi}$ denote the corresponding first order, second order inclusion probabilities, the weight and the vector of auxiliary variables for unit i in the h^{th} RHG.

Assuming $\Pr(m_h \leq 1)$ is negligible for $h = 1, . . ., H$, we have the following well known results for nonresponse adjustment using RHGs. In the sequel, "unbiased" is defined in the QR sense, that is, the QR expectation of the estimator equals to the quantity being estimated.

2.1. *Known Result 1 (Särndal et al. 1992, 581)*

The QR IPW estimator of T defined by $\hat{T}_1 = \sum\limits_{h=1}^{H} \sum\limits_{i \in s_h} \frac{\delta_{hi} d_{hi} y_{hi}}{\hat{\rho}_h}$ is asymptotically unbiased for T. In addition:

$$V\left(\hat{T}_1\right) = \sum_{i \in U} \sum_{j \in U} \Delta_{ij} \breve{y}_i \breve{y}_j + E_p E_m \left(\sum_{h=1}^{H} n_h^2 \frac{1 - \rho_h}{m_h} S_{1h}^2 | s \right) \triangleq V_1\left(\hat{T}_1\right) + V_2\left(\hat{T}_1\right)$$

where $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$, $\breve{y}_i = \frac{y_i}{\pi_i}$, $E_{\boldsymbol{m}}(.|s)$ is the expectation with respect to the sampling distribution of $\boldsymbol{m} = (m_1, .., m_h, . . ., m_H)^T$ given s, S_{1h}^2 is the variance of $\breve{y}_{hi}$ in s_h and $\breve{y}_{hi} = \frac{y_{hi}}{\pi_{hi}}$.

Finally, an approximately unbiased estimator of the variance, $V(\hat{T}_1)$, is given by:

$$\hat{V}(\hat{T}_1) = \sum_{i \in U} \sum_{j \in U} \frac{\Delta_{ij}}{\hat{\rho}_{ij}} \delta_i \breve{y}_i \delta_j \breve{y}_j + \sum_{h=1}^{H} n_h^2 \frac{1 - \hat{\rho}_h}{m_h} S_{1hr}^2$$

where S_{1hr}^2 is the variance of $\breve{y}_{hi}$ in S_{hr}, that is, the responding sample in s_h. In addition, $\hat{\rho}_{ij}$ is defined below.

(a) $\hat{\rho}_{ii} = \hat{\rho}_h = \dfrac{m_h}{n_h}$ for $i \in s_h$;

(b) $\hat{\rho}_{ij} = \dfrac{m_h}{n_h} \dfrac{m_h - 1}{n_h - 1}$ for i, $j \in s_h$ and $i \neq j$;

(c) $\hat{\rho}_{ij} = \dfrac{m_h}{n_h} \dfrac{m_{h'}}{n_h{}'}$ for $i \in s_h$ and $j \in s_{h'}$.

If the generalised regression estimator (GREG) is used instead of Horvitz-Thompson estimator in the QR IPW estimator of Result 1, we have the following result.

2.2. *Known Result 2 (Kim and Kim 2014; Särndal and Lundström 2005, 52; Särndal et al. 1992, 584)*

The QR IPW estimator of T defined by

$\hat{T}_2 = \sum_{h=1}^{H} \sum_{i \in s_h} \frac{\delta_{hi} d_{hi} g_{hi} y_{hi}}{\hat{\rho}_h}$, is asymptotically unbiased for T, where

$$g_{hi} = 1 + \left(\sum_{i \in U} \boldsymbol{x}_i - \sum_{h=1}^{H} \sum_{i \in s_h} \frac{\delta_{hi} d_{hi}}{\hat{\rho}_h} \boldsymbol{x}_{hi}\right)^T \left(\sum_{h=1}^{H} \sum_{i \in s_h} \frac{\delta_{hi} d_{hi}}{\hat{\rho}_h} c_{hi} \boldsymbol{x}_{hi} \boldsymbol{x}_{hi}^T\right)^{-1} (c_{hi} \boldsymbol{x}_{hi}) \text{ and } c_{hi}{}'s$$

are specified constants (often set as 1, see Särndal and Lundström 2005, 48) related to error structure of the estimation model underpinning GREG. Furthermore, analogous to Result 1, we have the following approximate variance:

$$V\left(\hat{T}_2\right) = \sum_{i \in U} \sum_{i \in U} \Delta_{ij} \breve{e}_i \breve{e}_j + E_p E_m \left(\sum_{h=1}^{H} n_h^2 \frac{1 - \rho_h}{m_h} S_{2h}^2 | s\right) \triangleq V_1\left(\hat{T}_2\right) + V_2\left(\hat{T}_2\right)$$

where $\breve{e}_i = \frac{\hat{e}_i}{\pi_i}$, $\hat{e}_i = y_i - \boldsymbol{x}_i^T \left(\sum_{h=1}^{H} \sum_{i \in s_h} \frac{\delta_{hi} d_{hi}}{\rho_{hi}} c_{hi} \boldsymbol{x}_{hi} \boldsymbol{x}_{hi}^T\right)^{-1} \left(\sum_{h=1}^{H} \sum_{i \in s_h} \frac{\delta_{hi} d_{hi}}{\rho_{hi}} c_{hi} \boldsymbol{x}_{hi} y_{hi}\right)$, S_{2h}^2 is the variance of $\breve{e}_{hi}$ in $s_{h,}$ $\breve{e}_{hi} = \frac{\hat{e}_{hi}}{\pi_{hi}}$, $\hat{e}_{hi} = y_{hi} - \boldsymbol{x}_{hi}^T \left(\sum_{h=1}^{H} \sum_{i \in s_h} \frac{\delta_{hi} d_{hi}}{\rho_{hi}} c_{hi} \boldsymbol{x}_{hi} \boldsymbol{x}_{hi}^T\right)^{-1} \left(\sum_{h=1}^{H} \sum_{i \in s_h} \frac{\delta_{hi} d_{hi}}{\rho_{hi}} c_{hi} \boldsymbol{x}_{hi} y_{hi}\right)$ Finally an approximately unbiased estimator of the variance, $V\left(\hat{T}_2\right)$, is given by:

$$\hat{V}\left(\hat{T}_2\right) = \sum_{i \in U} \sum_{j \in U} \frac{\Delta_{ij}}{\hat{\rho}_{ij}} \delta_i \breve{e}_i \delta_j \breve{e}_j + \sum_{h=1}^{H} n_h^2 \frac{1 - \hat{\rho}_h}{m_h} S_{2hr}^2$$

where $\hat{\rho}_{ij}$ is given in Result 1 and S_{2hr}^2 is the variance of $\breve{e}_{hi}$ in s_{hr} and

$$\hat{\breve{e}}_{hi} = \frac{y_{hi}}{\pi_{hi}} - \frac{\boldsymbol{x}_{hi}^T}{\pi_{hi}}\Big(\sum_{h=1}^{H}\sum_{i\in s_h}\frac{\delta_{hi}d_{hi}}{\hat{\rho}_h}c_{hi}\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T\Big)^{-1}\Big(\sum_{h=1}^{H}\sum_{i\in s_h}\frac{\delta_{hi}d_{hi}}{\hat{\rho}_h}c_{hi}\boldsymbol{x}_{hi}y_{hi}\Big)$$

If instead of weighting, imputation is used to mitigate the effect of nonresponse, we have the following result.

2.3. *Known Result 3 (Beaumont 2005)*

The QR Calibrated Imputation Estimator defined by $\hat{T}_3 = \sum_{i\in s} d_i g_i\{y_i + (1-\delta_i)y^*\}$ is asymptotically unbiased, where

$$y_i^* = \hat{\mu}_i + \frac{d_i g_i}{u_i}\left\{\sum_{i\in s}\frac{(1-\delta_i)\big(d_i g_i\big)^2}{u_i}\right\}^{-1}\left\{\sum_{h=1}^{H}\sum_{i\in s_h}\frac{1-\hat{\rho}_{hi}}{\hat{\rho}_{hi}}\delta_{hi}d_{hi}g_{hi}\big(y_{hi}-\hat{\mu}_{hi}\big)\right\},$$

$$\hat{\mu}_i = \boldsymbol{x}_i^T\Big(\sum_{h=1}^{H}\sum_{i\in s}\frac{\delta_{hi}c_{hi}\boldsymbol{d}_{hi}\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T}{\hat{\rho}_{hi}}\Big)^{-1}\Big(\sum_{h=1}^{H}\sum_{i\in s}\frac{\delta_{hi}c_{hi}\boldsymbol{d}_{hi}\boldsymbol{x}_{hi}y_{hi}}{\hat{\rho}_{hi}}\Big),$$

$$\hat{\mu}_{hi} = \boldsymbol{x}_{hi}^T\Big(\sum_{h=1}^{H}\sum_{i\in s}\frac{\delta_{hi}c_{hi}\boldsymbol{d}_{hi}\boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T}{\hat{\rho}_{hi}}\Big)^{-1}\Big(\sum_{h=1}^{H}\sum_{i\in s}\frac{\delta_{hi}c_{hi}\boldsymbol{d}_{hi}\boldsymbol{x}_{hi}y_{hi}}{\hat{\rho}_{hi}}\Big) \text{ and } u_i = \frac{1}{\hat{\sigma}_1^2} \text{ or } \frac{d_i g_i}{\hat{\sigma}_1^2}\,\hat{\sigma}_1^2$$

denotes the estimated variance of the linear model of y on $\boldsymbol{x}$. Furthermore, the variance of $\hat{T}_3$ is given by:

$$V\big(\hat{T}_3\big) = \sum_{i\in U}\sum_{j\in U}\Delta_{ij}\breve{e}_i\breve{e}_j + E_pE_m\Big(\sum_{h=1}^{H}n_h^2\frac{1-\rho_h}{m_h}S_{3h}^2|s\Big) \triangleq V_1\big(\hat{T}_3\big) + V_2\big(\hat{T}_3\big)$$

where $\breve{e}_i = \frac{y_i - \hat{\mu}_i}{\pi_i}$, S_{3h}^2 is the variance of $\breve{e}_{hi}$ in s_h and $\breve{e}_{hi} = \frac{y_{hi}-\hat{\mu}_{hi}}{\pi_{hi}}$ Finally, an asymptotically unbiased estimator of the variance, $V\big(\hat{T}_3\big)$, is given by:

$$\hat{V}\big(\hat{T}_3\big) = \sum_{i\in U}\sum_{j\in U}\frac{\Delta_{ij}}{\hat{\rho}_{ij}}\delta_i\breve{e}_i\delta_i\breve{e}_j + \sum_{h=1}^{H}n_h^2\frac{1-\hat{\rho}_h}{m_h}S_{3hr}^2$$

where $\hat{\rho}_{ij}$ is given in Result 1 and S_{3hr}^2 is the variance of $\breve{e}_{hi}$ in s_{hr}.

Note 1. The approach in Result 3 is described by Beaumont (2005) as calibrated imputation because $\hat{T}_3$ is calibrated to an asymptotically QR unbiased estimator of the population total based on $\hat{\mu}_i's$ (refer Equation 3.1 of Beaumont 2005). Also, some authors for example, Beaumont (2005) use the more theoretically correct expression, $\breve{e}_i^* = g_i\breve{e}_i$, in lieu of $\breve{e}_i$ in Results 2 and 3. However, Särndal and Lundström (2005, 37) pointed out in practice, $\breve{e}_i^* \simeq \breve{e}_i$, so numerically there is little difference in using $\breve{e}_i$ instead.

Note 2. Oh and Scheuren (1983) proposed the variance in Result 1 be conditional on $\boldsymbol{m}$. We shall follow this approach in the sequel. As a result, we can drop the $E_{\boldsymbol{m}}$ operator in $V_2\big(\hat{T}_j\big)$, $j = 1, 2, 3$ in Results 1 to 3 above and replace it by $V_2\big(\hat{T}_j|\boldsymbol{m}\big)$.

Note 3. From the similarity in the formulas for the variance of the three estimators in Results 1 to 3, we can develop a generic approach for the optimal allocation of follow-up resources in the next Section. The approach outlined in this article can be considered as the NFU resources allocation counterpart of Neyman allocation (Särndal et al. 1992, 106) for sampling resources.

Note 4. As mentioned earlier, if the propensity score model is missing not at random, that is, $\rho_i\ (\boldsymbol{x}_i,\ y_i;\boldsymbol{\alpha}) = \Pr(\delta_i = 1|\ \boldsymbol{x}_{i,}\ y_{i,}\ i \in s{:}\boldsymbol{\alpha})$, one can use $\frac{1}{\tilde{\rho}_i(\boldsymbol{x}_i|\boldsymbol{\alpha})}$ as the weight instead, where $\tilde{\rho}_i(\boldsymbol{x}_i|\boldsymbol{\alpha}) = \int \rho_i\ (\boldsymbol{x}_{i,}\ y_i\ |\ \boldsymbol{\alpha})\, f\ (y_i\ |\ \boldsymbol{x}_i)dy_i$ and Pfeffermann and Sverchkov (1999) showed that $\frac{1}{\tilde{\rho}_i(\boldsymbol{x}_i|\boldsymbol{\alpha})} = E\left(\frac{1}{\rho_i(\boldsymbol{x}_i,y_i;\boldsymbol{\alpha})}|\boldsymbol{x}_i, \delta_i = 1\right)$. By assuming the function form of $f\left(y_i|\boldsymbol{x}_i, \delta_i = 1; \boldsymbol{\beta}\right)$ one can use the respondent data to obtain a consistent estimate of $\boldsymbol{\beta}$ to compute $E\left(\frac{1}{\rho_i(\boldsymbol{x}_i,y_i;\boldsymbol{\alpha})}|\boldsymbol{x}_i, \delta_i = 1\right)$.

3. Optimum Allocation of Nonresponse Follow-Up Resources

Current practice of NFU in probability surveys is often based on an approach trying to convert as many nonrespondents as possible with a view to increasing the response rate. As argued in Curtin et al. (2000), Groves et al. (2004); Groves (2006); Groves and Peytcheva (2006), response rate is an insufficient indicator of survey data quality and an increase in response rate does not necessarily lead to a reduction in nonresponse bias. In this article, a strategic approach to NFU is proposed. Under this approach, the objective for NFU, under a fixed follow-up budget, is to minimise the mean squared error of the corresponding estimates. Given Results 1 to 3 above and note 2, this is equivalent to minimising $V_2\left(\hat{T}_j|\boldsymbol{m}\right)$, $j = 1, 2, 3$, that is, the nonresponse variance, as the sampling variance, $V_1\left(\hat{T}_j\right)$, is not numerically affected in a NFU setting.

We now model the k_h^{th} conversion event for the i^{th} nonrespondent in the h^{th} RHG as a random (Bernoulli) variable, λ_{khi}, with $\Pr(\lambda_{khi} = 1) = \rho_h$ and $\Pr(\lambda_{khi} = 0) = 1 - \rho_h$, where $\lambda_{khi} = 1$ if the conversion is successful. Note that, by the definition of RHGs, every nonrespondent in the same RHG has the same probability of a successful conversion. Moreover, we assume that the events λ_{khi} and λ_{lhj} are independent, that is, $\lambda_{khi} \perp \lambda_{lhj}$ where $k \neq l$ and $i \neq j$ and $h = 1, . . ., H; i,j \in U$. Whilst the assumptions $\lambda_{khi} \perp \lambda_{lhj}$, $i \neq j$ seem reasonable, particularly when the data collector for the i^{th} nonrespondent is different to that of the j^{th} nonrespondent, the assumptions $\lambda_{khi} \perp \lambda_{lhi}$, $k \neq l$ would hold only if the NFU visits are "passive" that is, inability to establish contacts in all previous visit until the visit resulting in a successful conversion. On the other hand, if the visits are non-passive, it is likely that these visits will impact the nonrespondent's decision to participate in the survey during the current visit, for example, persuasion effect as k grows, or alternatively, a "no" in the previous visit is the ultimate position of the nonrespondent. In this case, the assumption $\lambda_{khi} \perp \lambda_{lhi}$, $k \neq l$ is violated. Nevertheless, the Bernoulli assumption, as a working assumption, is helpful to guide the allocation of NFU resources and benign to the estimation process which uses, amongst other things, the actual number of nonrespondents converted in the estimation process, including the calculation of the estimated variance.

Under the Bernoulli model, the probability of conversion, after k_h visits for a nonrespondent in the h^{th} RHG, is $k_h\rho_h$. Here, we restrict $k_h \leq \frac{1}{\rho_h}$, as a nonrespondent cannot be converted more than once, that is, $k_h\rho_h > 1$ does not make sense. With $n_h - m_h$ nonrespondents in the h^{th} stratum, the expected number of additional respondents converted, after k_h attempts, is $k_h\rho_h(n_h - m_h)$. With respondents increased from m_h to $m_h + k_h\rho_h(n_h - m_h)$, whilst the sampling variance term, that is, $V_1\left(\hat{T}_j\right)$, of $V\left(\hat{T}_j\right), j = 1, 2, 3$ remains the same, we have:

$$V_2\left(\hat{T}_j|\boldsymbol{m}\right) = \sum_{h=1}^{H} n_h^2 \frac{1-\rho_h}{m_h + k_h\rho_h(n_h - m_h)} S_{jh}^2$$

$$\hat{V}_2\left(\hat{T}_j|\boldsymbol{m}\right) = \sum_{h=1}^{H} n_h^2 \frac{1-\hat{\rho}_h}{m_h + k_h\hat{\rho}_h(n_h - m_h)} S_{jhr}^2 \qquad (1)$$

where S_{jh}^2 and S_{jhr}^2 are defined in Results 1 to 3 respectively, depending on the type of weighting adjustment for nonresponse used.

On the other hand, the cost c_h of following up $n_h - m_h$ nonrespondents in the h^{th} RHG will be $u_h k_h(n_h - m_h)$, where u_h is the unit cost per visit to convert, and the cost of following up all nonrespondents across all RHGs is

$$C = \sum_{h=1}^{H} u_h k_h(n_h - m_h). \qquad (2)$$

Thus, the strategic approach to NFU becomes finding $k_h's$ such that Equation (1) is minimised subject to: (a) $C \le C_0$ for a pre-determined C_0, where C is defined (2); (b) $k_h \le \frac{1}{\hat{\rho}_h}$; (c) k_h is an integer; and (d) $k_h \le k_0$, where k_0 is another pre-determined constant to ensure reasonable respondent load. As there is no closed form for the solution, we have to find it using a solver. We used Excel Solver as illustrated in the next Section. It is good at finding solutions for problems with multiple inputs subject to multiple constraints. Note that the solutions are intuitively similar to the Neyman allocation for strata sample sizes–RHGs with high nonresponse variance is allocated with large follow-up resources, but tempered (in a non-linear way) by the unit cost of NFU effort.This is to boost the responding sample size needed to reduce the nonresponse variance for the relevant RHGs.

4. A Numerical Example

We tested the methods outlined in the previous Section through an empirical study, using data from the 2018/19 Rural Environment and Agricultural Commodities Survey (REACS) of the ABS (Australian Bureau of Statistics 2000). Conducted annually, the REACS releases statistics on the production of agricultural commodities including cereal and broadacre crops, fruit and vegetables and livestock on Australian farms. The REACS sample comprised a stratified simple random sample of Australian farms. Like many ABS probability surveys, REACS was confronted with the challenge of declining response rates. For this empirical study, the variable of interest, that is, the production of sheep, is illustrated. The sample for this variable of interest has 4,696 units (i.e., farms) in total, of which, 3,525 units were continuing units and 1,171 were new units, that is, units first rotated into the 2018/19 survey.

To predict the response propensity scores for continuing units, the Random Forest (RF) with regression trees algorithm (Breiman 2001) was used. The RF algorithm is used in this article instead of logistic regression models for predicting propensity scores for a number of reasons. For example, unlike logistic regression models, RF does not require the assumption of linearity or additivity in the modelling (Lee et al. 2010). Equally important, the automatic interaction detection inherent in RF algorithms provides a straightforward way to account for and allow easy interpretations of interactions between auxiliary data and the propensity to respond (Phipps and Toth 2014; Buskirk and Kolenikov 2015). For the RF algorithm used in this example, variables such as state, industry, size, statistical significance of the unit and such paradata as number of calls on the unit in previous survey cycles, number of reminder letters previously sent etc. were included in the model as predictors. Finally, the RF algorithm used in this example to predict propensity scores was chosen based on ten-fold cross validation (Hastie et al. 2009, 181), and has the smallest “out of sample” mis-classification rate in survey participation amongst the candidate RFs with different number of variable splits or trees.

For new units rotated into the 2018/19 REACS, the RF algorithm cannot be used to predict the propensity scores because para data from previous survey cycles for these units do not exist. Instead, their propensity scores were imputed by the average of the propensity scores of the continuing units considered to be their k Nearest Neighbours (kNN) (Hastie et al. 2009, 463). To run the kNN algorithm, four variables from the 2018/19 REACS including state, natural resources management region, industry and size were used to calculate the “distance” metric. Because they are categorical variables, the Gower’s Distance (Gower 1971) was used as the distance metric to find the nearest neighbours.

What value of k should be used? The optimal k should be one that gives the most accurate out-of-sample predictions. To find the optimal k, ten-fold cross validation was again applied to the (training) dataset comprising the estimated propensity scores of the continuing units. The root mean squared error of the predicted propensity scores was used to measure and compare the accuracy of twelve kNN models for k = 1,2, . . . ,12. As can be seen from Figure 1, whilst the 12NN model appears to be the most accurate, its RMSE is, however, not significantly different from that of the 5NN model. As the training dataset itself is skewed towards RHGs with high propensity scores, a small k is preferred, lest the KNN induced propensity scores of the new units are skewed by high propensity scores donors. Based on these considerations, the optimal k is taken to be five and the 5NN model was chosen to impute the propensity scores for the new units.

Finally, in REACS 2018/19, as the sample was stratified, and the stratified estimator was recast as a GREG estimator (Särndal and Lundström 2005, 37), so that the QR IPW estimator of Result 2, $\hat{T}_2$, is used in the numerical example.

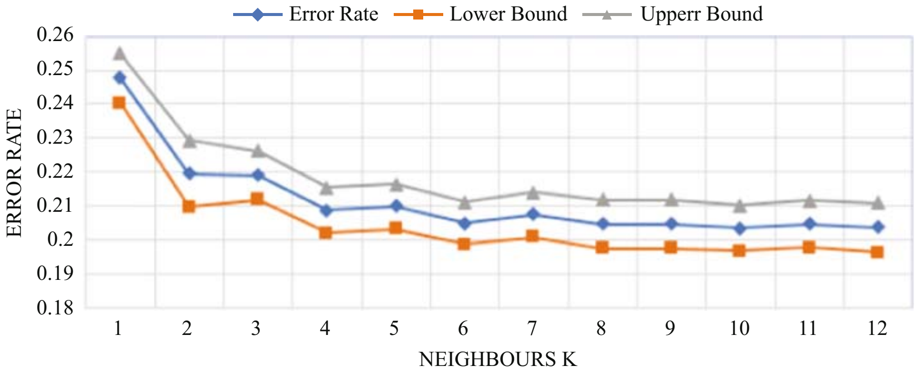


Fig. 1. Estimated out-of-sample error rate by different value of k in kNN models.

Table 1 gives the relevant information with respect to the estimator of Result 2. In this example, we set $C_0 = \text{USD } 30,000$ and $k_0 = 6$. We used the complimentary Excel Solver to solve the problem by creating an objective "cell" in Excel (i.e., Equation (1)), a 10x1 vector of k_h input "cells", three 10x1 vector of constraint "cells" (i.e., constraints (b) to (d) in Section 3 above), and a total cost constraint "cell" (i.e., constraint (a)). When running the Solver, Excel allows the 10x1 vector of k_h input "cells" to vary, subject to meeting the requirements of the constraint cells, in order to find the solutions to minimise the objective cell. Table 1 gives the relevant data to compute the optimal solutions, and $\hat{V}_2\left(\hat{T}_2|\mathbf{m}\right)$. The optimum solutions are given in Table 2.

In this example, the investment of about USD 19.6k (see Table 2) for NFU is cost effective, as it reduces the nonresponse variance, $\hat{V}_2\left(\hat{T}_2|\mathbf{m}\right)$, from 57.77 billion units (Table 1) where there is no NFU to about 38.21 billion units (Table 2) by restricting the number of follow-up

Table 1. REACS 2018/19 information on $\hat{T}_2$ for sheep and $\hat{V}_2\left(\hat{T}_2|m\right)$ with no NFU.

RHG	Propensity	n_h	$\hat{\rho}_h$	S^2_{jhr}	m_h	u_h	$\hat{V}_2(\hat{T}_2\mid\mathbf{m})$
	score			(10^7)		(USD)	(10^9)
	range						
1	(0,0.1]	23	0.09	2.30	2	20	5.55
2	(0.1,0.2]	23	0.17	3.32	4	20	3.64
3	(0.2,0.3]	37	0.27	1.58	10	20	1.58
4	(0.3,0.4]	79	0.35	2.50	28	20	3.60
5	(0.4,0.5]	117	0.45	2.72	53	20	3.84
6	(0.5,0.6]	178	0.55	4.36	98	20	6.33
7	(0.6,0.7]	285	0.65	8.49	186	20	12.87
8	(0.7,0.8]	257	0.75	6.51	193	20	5.55
9	(0.8,0.9]	504	0.85	7.90	429	20	6.96
10	(0.9,1)	3193	0.95	4.69	3034	20	7.85
Total		4696			4037		57.77

Table 2. Optimal k_h and $\hat{V}_2\left(\hat{T}_2|\mathbf{m}\right)$ for C_0 – USD 30k and $k_0 = 6$

RHG	RP	k_h	Additional	RHG follow	$\hat{V}_2(\hat{T}_2\mid\mathbf{m})$
	score		respondents* –	up cost (USD)	(10^9)
	range		$k_h\hat{\rho}_h(n_h - m_h)$	$u_hk_h(n_h - m_h)$	
1	(0,0.1]	6	11	2,520	0.86
2	(0.1,0.2]	5	17	1,900	0.93
3	(0.2,0.3]	2	15	1,080	0.70
4	(0.3,0.4]	2	36	2,040	1.75
5	(0.4,0.5]	2	58	2,560	2.08
6	(0.5,0.6]	1	44	1,600	4.37
7	(0.6,0.7]	1	64	1,980	9.56
8	(0.7,0.8]	1	48	1,280	4.44
9	(0.8,0.9]	1	64	1,500	6.06
10	(0.9,1)	1	151	3,180	7.45
Total	-	-	508	19,640	38.21

*Rounded to the nearest integer

visits to the values of k_h as shown in Table 2. We can also see from Table 2 that there is an increase of 508 respondents with the optimal allocation of NFU resources, lifting the pre-follow up response rate of 86.0% to a post- follow up response rate of 97%. On the other hand, using the same Excel Solver, we can show that if only, say, USD 10k is available for follow-up, the optimal $k_h's$ are $k_1 = 2, k_2 = 3, k_9 = k_{10} = 0$, and $k_i = 1$, for $i = 3, 4, 5, 6, 7, 8$. This would incur a total investment of USD 9.6k; gives a nonresponse variance of 41.9 billion units, and an increase of 225 respondents with a post-follow up response rate of only 91%.

With a sampling variance of 41.3x10^9 units for sheep production, the use of optimal NFU allocation reduces the mean squared error from 99.11x10^9 (with no NFU follow-up) to 79.55x10^9 units, or a reduction by 19.7% and 10.4% in the mean squared error and root mean squared error respectively.

It is also instructive to compare the optimal allocation strategy with the common practice, where there is no differentiation in the number of visits across the RHGs, that is, $k_h = k_0$, $h = 1, \ldots, 10$. From (2), $k_0 = C_o / \sum_{h=1}^{H} u_h(n_h - m_h) =$ USD $30,000/\{$USD $20*(4,696-4,037)\} \simeq 2$. As compared with the optimal allocation, we see that the common practice in this example does not put enough resources for the first two RHGs, and too much for the last five RHGs. The result is that the nonresponse variance of 39.84 billion units under the constant k_0 approach is higher than the optimum nonresponse variance of 38.21 billion units. Even though the reduction in mean squared error is 3.9%, the common practice approach costs USD 10.4k more than the optimal NFU allocation approach of USD 19.6k. However, with the number of additional respondents at 596 under the common approach, it gives a response rate of 99%. This is an example showing the response rate of 99% can give a misleading impression of providing a higher quality estimate of the number of sheep produced than the one with a response rate 97% with optimal NFU allocation.

If one is interested in looking at the return on the investment of NFU resources, as measured by the reduction of $\hat{V}_2(\hat{T}_2|\boldsymbol{m})$ per unit NFU cost, one can use different values of C_0 to provide the requisite information as shown in Table 3. We observe that the highest return on the NFU investment is to spend about USD 10k for NFU, where $1,091$x10^3 units of variance is reduced per unit cost. However, the strategy to choose the highest return on NFU resources for NFU is only sensible if the nonresponse variance of 41.91 billion units is acceptable.

Table 3. Optimal $\hat{V}_2(\hat{T}_2|\boldsymbol{m})$ and Response Rate for $k_0 = 6$ and different values of C_0

Cost (USD) (not greater than)	$\hat{V}_2(\hat{T}_2\|\boldsymbol{m})$ (10^9)	Response rate (%)	Actual cost (USD)	Reduction in variance (10^3) per additional unit cost
5,000	47.26	88	4,600	-
7,500	44.42	89	7,380	1,022
10,000	41.91	91	9,680	1,091
12,500	40.56	92	12,200	536
15,000	39.24	93	14,900	489
17,500	39.02	96	17,500	85
20,000	38.21	97	19,640	379
20,250	38.21	97	19,640	-

5. Conclusion

In this article, we provide a method to allocate NFU resources in such a way as to minimise the nonresponse variance (and thus the MSE) of the estimator of a variable of interest in probability surveys. This method requires the use of a quasi-randomisation framework and an inverse propensity weight to adjust for nonresponse.

In the numerical example, we used Random Forest to estimate the response propensity scores for the continuing units and applied the kNN algorithm to impute the propensity scores for the new units first rotated into the 2018/19 REACS. In both cases, selection of the RF algorithm, or the value of k in the kNN algorithm, was based on 10 fold cross validation, and the RF algorithm/k in the kNN algorithm with the least out-of-sample classification/prediction error was chosen. When compared with the common practice, the numerical example shows that there is a further reduction (about 4%) in the nonresponse variance $\hat{V}_2\left(\hat{T}_2|\boldsymbol{m}\right)$ by adopting the optimum allocation of NFU resources with only two third of the cost. The numerical example also shows that response rate is an inadequate indicator of data quality and the common NFU practice is not the most cost effective.

We note that official surveys are multi-purpose and optimising the allocation of NFU resources with respect to one target variable may not result in the optimal allocation of this resources to the other variables. This issue is also faced by survey statisticians in using Neyman allocation. An approach commonly used in official surveys to address this issue is to allocate the sample sizes with reference to the highest priority variable. Another way is to define objective functions for optimisation which address the multivariate situation. This is discussed in Holmberg (2002, 2003) and Holmberg et al. (2003). The same approaches can be used to allocate NFU resources in multiple purpose surveys.

The important messages of this article are (1) reinforcing the well-known fact that allocating significant resources to boost response rates will not necessarily lead to better quality statistics; (2) in order to achieve better statistical outcomes, the NFU resource allocation should instead aim at minimising the variance from nonresponse; and (3) optimal allocation of NFU resources will result in the most cost effective NFU practice.

6. References

Australian Bureau of Statistics, 2000. *Droughts, fires, cyclones, hailstorms and a pandemic*. Available at: https://www.abs.gov.au/articles/droughts-fires-cyclones-hailstorms-and-pandemic-march-quarter-2020#businesses.

Beaumont, J.F., 2005. "Calibrated Imputation In Surveys Under A Quasi-Model-Assisted Approach." *Journal of the Royal Statistical Society* B67: 445–458. DOI: https://doi.org/10.1111/j.1467-9868.2005.00511.x.

Beaumont, J.F., C. Bocci, and D. Haziza. 2014. "An Adaptive Data Collection Procedure for CallPrioritization." *Journal of Official Statistics* 30: 607–621. DOI: http://dx.doi.org/10.2478/jos-2014-0040.

Bethlehem, J.G. 1998. "Reduction of Nonresponse Bias Through Regression Estimation." *Journal of Official Statistics* 3: 251–260.

Breiman, L. 2001. "Random Forests." *Machine Learning* 45: 5–32. DOI: https://doi.org/10.1023/A:1010933404324.

Buskirk, T.D., and S. Kolenikov. 2015. "Finding Respondents in the Forest: A Comparison of Logistic Regression and Random Forest Models for Response Propensity Weighting and Stratification." *Survey Methods: from the Field*. DOI: http://doi.org/10.13094/SMIF-2015-00003.

Curtin, R., S. Presser, and E. Singer. 2000. "The Effects of Response Rate Changes on the Index of Consumer Sentiment." *Public Opinion Quarterly* 64: 413–428. DOI: http://dx.doi.org/10.1086/318638.

Elliott, M.R., R.J.A. Little, and S. Lewitzky. 2000 "Subsampling Callbacks to Improve Survey Efficiency." *Journal of the American Statistical Association* 95: 730–738. DOI: https://dx.doi.org/10.2307/2669453.

Gower, J.C. 1971. "A Coefficient of Similarity and Some of Its Properties." *Biometrics* 27: 857–874. DOI: https://dx.doi.org/10.2307/2528823.

Groves, R.M., 2006 "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70: 646–675. DOI: https://dx.doi.org/10.1093/poq/nfl033.

Groves, R.M., and M.P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: John Wiley and Sons.

Groves, R.M., and S.G. Heeringa. 2006. "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society* 169: 439–457. DOI: https://doi.org/10.1111/j.1467-985X.2006.00423.x.

Groves, R.M., and E. Peytcheva. 2006. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." In Proceedings of the 17th International Workshop on Household Survey Nonresponse, August 28–30, Omaha, NE, USA. DOI: https://doi.org/10.1093/poq/nfn011.

Groves, R.M., S. Presser, and S. Dipko. 2004. "The Role of Topic Interest in Survey Participation Decisions." *Public Opinion Quarterly* 68: 2–31. DOI: http://dx.doi.org/10.1093/poq/nfh002.

Hansen, M.H., and W.N. Hurwitz. 1946. "The Problem of Nonresponse in Sample Surveys." *Journal of the American Statistical Association* 41: 517–529.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Second Edition, New Youk: Springer.

Hedlin D. 2020. "Is There A 'Safe Area' Where the Nonresponse Rate Has Only A Modest Effect on Bias Despite Non-Ignorable Nonresponse?" *International Statistical Review* 88: 642–657. DOI:. https://dx.doi.org/10.1111/insr.12359

Holmberg, A. 2002. "A Multiparameter Perspective on the Choice of Sampling Design in Surveys." *Statistics in Transition* 6: 969–994.

Holmberg, A. 2003. "Using Auxiliary Information to Choose between Alternative Sampling Designs in A Survey with Several Key Variables." In Proceedings of the Section on Challenges in Survey Taking for the Next Decade: Statistics Canada Symposium, October 29, Quebec: Available at: https://www150.statcan.gc.ca/n1/en/pub/11-522-x/2003001/session18/7726-eng.pdf?st = FCpnf4OS. (accessed February 2023).

Holmberg, A., P. Flisberg, and M. Rönnqvist. 2003. *On the choice of optimal design in business surveys with several important study variables*. Uppsala: Uppsala University. Available at: https://www.diva-portal.org/smash/get/diva2:162729/FULLTEXT01.pdf (accessed February 2023).

Kim, J.K., and J.J. Kim. 2014. "Nonresponse Weighting Adjustment Using Estimated Response Probability." *The Canadian Journal of Statistics* 35:501–514. DOI: https://doi.org/10.1002/cjs.5550350403.

Kim, J.K., and K. Morikawa. 2022. *An Empirical Likelihood Approach to Reduce Selection Bias in Voluntary Samples*. New York: Cornell University. Available at: https://arxiv.org/abs/2211.02998.

Kim, J.K., and J. Shao. 2014. *Statistical Methods for Handling Incomplete Data*. London: Chapman and Hall.

Lee, B.K., J. Lessler, and E.A. Stuart. 2010. "Improving Propensity Score Weighting Using Machine Learning." *Statistics in Medicine* 29: 337–346. DOI: https://dx.doi.org/10.1002/sim.3782.

Little, R.J.A. 1986. "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review* 54: 139–157. DOI: https://dx.doi.org/10.2307/1403140.

Little, R.J.A., and D.B. Rubin. 2019. *Statistical Analysis of Missing Data*. New York: John Wiley and Sons.

Neusy, E., J.F. Beaumont, W. Yung, M. Hidiroglou, and D. Haziza 2022. "Nonresponse Follow-Up for Business Surveys." *Survey Methodology* 48: 95–117. Available at: http://www.statcan.gc.ca/pub/12-001-x/2022001/article/00006-eng.htm.

Oh, H.L., and F.J. Scheuren. 1983 "Weighting Adjustment for Unit Nonresponse." *Incomplete Data in Sample Surveys*, 2: 143–184.

Pfeffermann, D., and M. Sverchkov. 1999. "Parametric and Semiparametric Estimation of Regression Models Fitted to Survey Data." *The Indian Journal of Statistics* 61: 166–186. Available at: https://www.jstor.org/stable/i25053064.

Phipps, P., and D. Toth 2014. *Regression Tree Models for Analyzing Survey Response*. U.S. Bureau of Labour Statistics. Washington DC: BLS. Available at: https://www.bls.gov/osmr/research-articles/2014/ pdf/st140160.pdf (accessed February 2023).

Särndal, C.E., B. Swensson, and J. Wretman 1992. *Model Assisted Sampling*. New York: Springer.

Särndal, C.E., and S. Lundström 2005. *Estimation in Surveys with Nonresponse*. New York: John Wiley and Sons.

Schouten, B., F. Cobben, and J. Bethlehem. 2009. "Indicators for the Representativeness of Survey Response." *Survey Methodology* 35: 101–113.

Schouten, B., F. Cobben, and J. Bethlehem. 2011. *Handbook of Nonresponse in Household Surveys*. New York: John Wiley and Sons.

Sikov, A. 2018. "A Brief Review of Approaches to Non-ignorable Nonresponse: Approaches to Non-ignorable Nonresponse." *International Statistical Review* 86: 415–441. DOI: https://doi.org/10.1111/insr.12264.

Received August 2022
Revised December 2022
Accepted February 2023